

Christina Sanchez-Stockhammer

English Compounds and their Spelling

ENGLISH COMPOUNDS AND THEIR SPELLING

Anyone writing texts in English is constantly faced with the unavoidable question whether to use open spelling (*drinking fountain*), hyphenation (*far-off*) or solid spelling (*airport*) for individual compounds. While some compounds commonly occur with alternative spellings, others show a very clear bias for one form. This book tests more than sixty hypotheses and explores the patterns underlying the spelling of English compounds from a variety of perspectives. Based on a sample of 600 biconstituent compounds with identical spelling in all reference works in which they occur (200 each with open, hyphenated and solid spelling), this empirical study analyses large amounts of data from corpora and dictionaries and concludes that the spelling of English compounds is not chaotic but actually correlates with a large number of statistically significant variables. An easily applicable decision tree is derived from the data and an innovative multidimensional prototype model is suggested to account for the results.

Christina Sanchez-Stockhammer is a senior lecturer in English linguistics at Ludwig-Maximilian University of Munich. She has published widely on many diverse topics. Her books include *Consociation and Dissociation: An Empirical Study of Word-Family Integration in English and German* (2008), *Can We Predict Linguistic Change?* (2015, as editor) and (as co-editor) *Variational Text Linguistics: Revisiting Register in English* (2016).

STUDIES IN ENGLISH LANGUAGE

General editor

Merja Kytö (Uppsala University)

Editorial Board

Bas Aarts (University College London)

John Algeo (University of Georgia)

Susan Fitzmaurice (University of Sheffield)

Christian Mair (University of Freiburg)

Charles F. Meyer (University of Massachusetts)

The aim of this series is to provide a framework for original studies of English, both present-day and past. All books are based securely on empirical research, and represent theoretical and descriptive contributions to our knowledge of national and international varieties of English, both written and spoken. The series covers a broad range of topics and approaches, including syntax, phonology, grammar, vocabulary, discourse, pragmatics and sociolinguistics, and is aimed at an international readership.

Already published in this series

Christiane Meierkord: *Interactions across Englishes: Linguistic Choices in Local and International Contact Situations*

Haruko Momma: *From Philology to English Studies: Language and Culture in the Nineteenth Century*

Raymond Hickey (ed.): *Standards of English: Codified Varieties around the World*

Benedikt Szmrecsanyi: *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*

Daniel Schreier and Marianne Hundt (eds.): *English as a Contact Language*

Bas Aarts, Joanne Close, Geoffrey Leech and Sean Wallis (eds.): *The Verb Phrase in English: Investigating Recent Language Change with Corpora*

Martin Hilpert: *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*

Jakob R. E. Leimgruber: *Singapore English: Structure, Variation and Usage*

Christoph Rühlemann: *Narrative in English Conversation*

Dagmar Deuber: *English in the Caribbean: Variation, Style and Standards in Jamaica and Trinidad*

Eva Berlage: *Noun Phrase Complexity in English*

Nicole Dehé: *Parentheticals in Spoken English: The Syntax–Prosody Relation*

Jock Onn Wong: *English in Singapore: A Cultural Analysis*

Anita Auer, Daniel Schreier and Richard J. Watts: *Letter Writing and Language Change*

- Marianne Hundt: *Late Modern English Syntax*
- Irma Taavitsainen, Merja Kytö, Claudia Claridge and Jeremy Smith: *Developments in English: Expanding Electronic Evidence*
- Arne Lohmann: *English Co-ordinate Constructions: A Processing Perspective on Constituent Order*
- John Flowerdew and Richard W. Forest: *Signalling Nouns in English: A Corpus-Based Discourse Approach*
- Jeffrey P. Williams, Edgar W. Schneider, Peter Trudgill and Daniel Schreier: *Further Studies in the Lesser-Known Varieties of English*
- Nuria Yáñez-Bouza: *Grammar, Rhetoric and Usage in English: Preposition Placement 1500–1900*
- Jack Grieve: *Regional Variation in Written American English*
- Douglas Biber and Bethany Gray: *Grammatical Complexity in Academic English: Linguistics Change in Writing*
- Gjertrud Flermoen Stenbrenden: *Long-Vowel Shifts in English, c. 1050–1700: Evidence from Spelling*
- Zoya G. Proshina and Anna A. Eddy: *Russian English: History, Functions, and Features*
- Raymond Hickey: *Listening to the Past: Audio Records of Accents of English*
- Phillip Wallage: *Negation in Early English: Grammatical and Functional Change*
- Marianne Hundt, Sandra Mollin and Simone E. Pfenninger: *The Changing English Language: Psycholinguistic Perspectives*
- Joanna Kopaczyk and Hans Sauer (eds.): *Binomials in the History of English: Fixed and Flexible*
- Alexander Haselow: *Spontaneous Spoken English*
- Christina Sanchez-Stockhammer: *English Compounds and Their Spelling*
- Earlier titles not listed are also available*

ENGLISH COMPOUNDS AND THEIR SPELLING

CHRISTINA SANCHEZ-STOCKHAMMER

Ludwig-Maximilian University of Munich



CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107197848

DOI: 10.1017/9781108181877

© Christina Sanchez-Stockhammer 2018

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2018

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library.

ISBN 978-1-107-19784-8 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

*I dedicate this book to my husband, Philipp,
our children, Sophie and Julius,
and my uncle Roland*

Contents

<i>List of Figures</i>	<i>page</i> xi
<i>List of Tables</i>	xii
<i>Acknowledgements</i>	xx
1 Introduction 1	
1.1 The State of the Art in English Compound Spelling	4
1.2 Summary and Aims of the Present Study	17
PART I THEORETICAL BACKGROUND 21	
2 Delimitating the Compound Concept 23	
2.1 Compounds versus Phrases	24
2.2 Compounds versus Other Lexemes	35
2.3 Compounds versus Multi-word Items	39
2.4 Compounds versus Names	42
2.5 Compound Types	43
2.6 Summary	57
3 The Normative Background 61	
3.1 Why 'Correct' Spelling Matters	62
3.2 The Originators of Spelling Norms	66
3.3 Summary	73
PART II EMPIRICAL STUDY OF ENGLISH COMPOUND SPELLING 75	
4 Material and Method 77	
4.1 Dictionaries	77
4.2 Corpora	81
4.3 CompSpell	88
5 Potential Determinants of English Compound Spelling 96	
5.1 Spelling	97
5.2 Length	112

5.3	Frequency	132
5.4	Phonology	145
5.5	Morphology	153
5.6	Grammar	170
5.7	Semantics	197
5.8	Diachronic Variables	210
5.9	Discourse Variables	226
5.10	Systemic Variables	233
5.11	Extralinguistic Variables	244
5.12	General Issues	254
5.13	Summary	264
PART III MODELLING ENGLISH COMPOUND SPELLING		273
6	Compound Spelling Heuristics	275
6.1	Method	276
6.2	Results	281
6.3	Discussion	295
6.4	Summary	304
7	Modelling English Compound Spelling	307
7.1	The Relation between the Three Main Spelling Variants	307
7.2	Prototype Theory	313
7.3	Analogy	324
7.4	Cognitive Perspectives on English Compound Spelling	328
7.5	Integrating Change	339
7.6	Summary	344
8	Summary and Conclusion	347
	<i>Appendix</i>	356
	<i>References</i>	375
	<i>Index</i>	394

Figures

6.1	Decision tree for the comprehensive Algorithm 1 (all significant variables) for OHS_600	<i>page</i> 283
6.2	Decision tree for the maximally efficient Algorithm 4 for OHS_600	288
7.1	The relationship between open, hyphenated and solid spelling	311
7.2	Preliminary prototype structure of the category ENGLISH COMPOUNDS WITH SOLID SPELLING	315
7.3	Modified prototype structure of the category ENGLISH COMPOUNDS WITH SOLID SPELLING	316
7.4	Overlapping prototype model for compounds with open, hyphenated and solid spelling	318
7.5	Modified overlapping prototype model	320
7.6	A three-dimensional prototype model of English compound spelling	321

Tables

2.1	Order of English adjectives within the noun phrase	<i>page</i> 30
2.2	Compound types based on word formation	45
2.3	Labels used for compound classification in Table 2.4	48
2.4	Compound types based on part of speech	49
2.5	The delimitation of compounds	59
4.1	Dictionary lemma lists used in the empirical study	79
4.2	Corpora used in the empirical study	81
4.3	Make-up of the CompText corpus	86
4.4	CompSpell inflection tolerance principles (= part-of-speech-specific inflection which may be added to the compounds in the corpus searches)	91
5.1	Grouped consonant clusters across the constituent joint [o_1_CC_2_r] and spelling in OHS_600	100
5.2	Identical graphemes [Ident_lett] across the constituent joint in OHS_600	101
5.3	Grouped identical graphemes across the constituent joint [Ident_lett_r] and spelling in OHS_600	102
5.4	Garden path clusters across the constituent joint [Digraph] in OHS_600 and OHS_extra	104
5.5	Grouped garden path clusters across the constituent joint [Digraph_r] and spelling in OHS_extra	105
5.6	Grouped vowel graphemes across the constituent joint [o_1_VV_2_r] and spelling in OHS_600	106
5.7	Apostrophes [Apostr] and spelling in OHS_extra	110
5.8	Master_5+ compounds with more than two constituents which contain apostrophes but no genitive	111
5.9	CompSpell's syllabification principles	115
5.10	Number of syllables of the whole compound [Syll_total] in OHS_600	116

5.11	Grouped number of syllables of the whole compound [Syll_total_r] and spelling in OHS_600	117
5.12	Number of letters of the whole compound [Lett_total] in OHS_600	118
5.13	Grouped number of letters of the whole compound [Lett_total_r] and spelling in OHS_600	119
5.14	Compounds whose constituents exceed a ratio of 2:1/1:2 (in number of letters) [Lett_diff_12] and spelling in OHS_600	121
5.15	Compounds whose constituents exceed a ratio of 2:1/1:2 (in number of syllables) [Syll_diff_12] and spelling in OHS_600	123
5.16	Grouped number of letters of first constituent [Lett_1_r] and spelling in OHS_600	125
5.17	Grouped number of letters of second constituent [Lett_2_r] and spelling in OHS_600	126
5.18	Grouped length of first constituent (in syllables) [Syll_1_r] and spelling in OHS_600	128
5.19	Grouped length of second constituent (in syllables) [Syll_2_r] and spelling in OHS_600	129
5.20	Grouped unlemmatised spelling-insensitive frequency in BNCwritten [Total_BNC_r] for the OHS_600 compounds	133
5.21	Grouped unlemmatised spelling-insensitive frequency in BNCwritten [Total_BNC_r] and spelling in OHS_600	134
5.22	Grouped unlemmatised spelling-insensitive frequency in BNCwritten [Total_BNC_r] and spelling of the noun+noun=noun compounds in OHS_600	135
5.23	Frequency ranges for the first and second constituents in OHS_600	136
5.24	Combined frequency ranges of the first and second constituents [Freq_rvs2] and spelling in OHS_600 (l = low, m = mid, h = high)	137
5.25	Grouped frequency difference between the constituents [Freq_diff_12_r] and spelling in OHS_600	140
5.26	Grouped frequency range of the first constituent [Freq_1_r] and spelling in OHS_600	141
5.27	Grouped frequency range of the second constituent [Freq_2_r] and spelling in OHS_600	142
5.28	Corpus frequency and spelling variation in the dictionaries for the Master List compounds	143

5.29	Possible combinations of stress and spelling in compounds based on the examples in Bauer (1983: 104 and 1998: 79)	146
5.30	Main stress [Stress] and spelling in OHS_600	149
5.31	Main stress [Stress] and spelling for the noun compounds in OHS_600	150
5.32	Main stress [Stress] and spelling for the adjective compounds in OHS_600	151
5.33	Non-compound-final lexical suffixes [Nonfin_lex_suff] in OHS_600	154
5.34	Grouped non-compound-final lexical suffixes [Nonfin_lex_suff_r] and spelling in OHS_600	155
5.35	Non-compound-final grammatical suffixes [Nonfin_gr_suff] in OHS_600	156
5.36	Grouped non-compound-final grammatical suffixes [Nonfin_gr_suff_r] and spelling in OHS_600	157
5.37	Non-compound-initial prefixes [Nonini_pref] in OHS_600	158
5.38	Compound-final <i>-ing</i> , <i>-ed</i> , <i>-er</i> and variants [Final_ingeder] in OHS_600	159
5.39	Grouped compound-final <i>-ing</i> , <i>-ed</i> or <i>-er</i> [Final_ingeder_r] and spelling in OHS_600	160
5.40	Complex constituents [Complex_const] in OHS_600	163
5.41	Grouped complex constituents [Complex_const_r] and spelling in OHS_600	164
5.42	Morphological structure [Morphol_struct] in OHS_600	166
5.43	Grouped morphological structure [Morphol_struct_r] in OHS_600	167
5.44	Morphological structure [Morphol_struct] and spelling in OHS_600	169
5.45	Part of speech of the whole compound [PoS_comp] in the Master List	172
5.46	Grouped part of speech of the whole compound [PoS_comp_r] and spelling in OHS_600	173
5.47	Adjective compounds with an initial <i>-ly</i> adverb in Master_I-4	175
5.48	Spelling of the adverb compounds in Master_5+	177
5.49	Part of speech of the first constituent [PoS_1] and spelling in OHS_600	181
5.50	Grouped part of speech of the second constituent [PoS_2_r] and spelling in OHS_600	182
5.51	Part-of-speech combinations for the first and second constituents in OHS_600 sorted by spelling variant	183

5.52	Predicted spelling preferences of the OHS_600 compounds based on the part of speech of the first constituent [PoS_1] and the part of speech of the second constituent [PoS_2]	184
5.53	Part of speech of the first constituent [PoS_1] and the second constituent [PoS_2] of the OHS_600 compounds with grouping as lexical/grammatical	185
5.54	Combination of lexical and grammatical constituents [Lexgr_12_diff] in OHS_600	186
5.55	Preferred spelling in attributive position in BNCwritten [Attr] and spelling in OHS_600	188
5.56	Preferred spelling in predicative position in BNCwritten [Pred] and spelling in OHS_600	190
5.57	Preferred spelling in non-attributive position in BNCwritten [Nonattr] and spelling in OHS_600	192
5.58	Spelling of the OHS_600 compounds favoured in attributive, predicative and non-attributive position in BNCwritten	192
5.59	Sentential paraphrases of the compound <i>bird+cage</i> (following Stein 1974: 321)	193
5.60	Syntactic structures of compounds following Quirk et al. (1985: 1570–1578)	195
5.61	Grouped general reference nouns [General_n_r] and spelling in OHS_600	199
5.62	General reference nouns [General_n] and spelling in OHS_extra	199
5.63	Compound-final general reference nouns [General_n] and spelling in OHS_extra	200
5.64	Semantic relations between compound constituents	201
5.65	Special semantic relations between the constituents [Sem_relation] in OHS_600	203
5.66	Special semantic relations between the constituents [Sem_relation] and spelling in OHS_600	204
5.67	Idiomatcity [Idiom] and spelling in OHS_600	208
5.68	Etymological origin of the constituents [Etym_orig] of the OHS_600 compounds	212
5.69	Mixed Germanic and Romance origin [Mixed_etym] and spelling in OHS_600	213
5.70	Germanic/Romance/mixed origin of constituents, average length and spelling in OHS_600	213

5.71	Synchronically felt foreignness [Foreignness] and spelling in OHS_600	215
5.72	Date of first attestation of the OHS_600 compounds in the OED [Age] by century	217
5.73	Grouped age of the compound [Age_r] and spelling in OHS_600	218
5.74	Spelling development of the OHS_600 compounds from their first attestation in the OED to their unanimous present-day dictionary spelling	220
5.75	Grouped earliest spelling in the OED [Earl_spell_r] and spelling in OHS_600	222
5.76	Spelling of the OHS_600 compound types in chronologically ordered British English corpora	223
5.77	Spelling of the OHS_600 compound types in chronologically ordered British English corpora, combined with the unanimous dictionary spelling	224
5.78	Spelling of <i>to+day</i> in chronologically ordered British English corpora	225
5.79	Spelling of the OHS_600 compound types in corpora of edited versus unedited texts, combined with the usual dictionary spelling	229
5.80	Spelling of the OHS_600 compound types in British versus American English corpora	230
5.81	Spelling of the OHS_600 compound tokens in British versus American English corpora	231
5.82	Spelling of the compound types in British versus American English corpora for the Master List compounds with only one occurrence in the dictionaries	232
5.83	Spelling of the compound tokens in British versus American English corpora for the Master List compounds with only one occurrence in the dictionaries	232
5.84	Spelling with the highest type frequency in the left constituent family [LS_r] and spelling in OHS_600	237
5.85	Spelling with the highest type frequency in the right constituent family [RS_r] and spelling in OHS_600	238
5.86	Spelling with the highest token frequency in the left constituent family [LF_r] and spelling in OHS_600	240
5.87	Spelling with the highest token frequency in the right constituent family [RF_r] and spelling in OHS_600	242

5.88	OHS_600 spellings predicted correctly by the constituent family variables	243
5.89	Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the NPS Chat Corpus	248
5.90	Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the NPS Chat Corpus, combined with the usual dictionary spelling	248
5.91	Spelling of the OHS_600 compound tokens in the Blog Authorship Corpus and the NPS Chat Corpus	249
5.92	Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the CorTxt Corpus	250
5.93	Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the CorTxt Corpus, combined with the unanimous dictionary spelling	251
5.94	Spelling of the OHS_600 compound tokens in the Blog Authorship Corpus and the CorTxt Corpus, combined with the unanimous dictionary spelling	252
5.95	Spelling of the OHS_600 compound tokens in the Blog Authorship Corpus and the CorTxt Corpus	252
5.96	Lexicalisation and spelling	259
5.97	General spelling tendencies extracted from Table A.9	262
5.98	Variables coded in the database	265
5.99	Extract from Table A.9	269
6.1	The comprehensive Algorithm 1 comprising all significant variables (initial version)	278
6.2	Predictive accuracy of the comprehensive Algorithm 1 (all significant variables) for OHS_600	282
6.3	Variables retained in the final version of the comprehensive Algorithm 1	282
6.4	Correlation between compound length (in syllables) and frequency (in BNCwritten) for OHS_600	284
6.5	The user-friendly Algorithm 2	285
6.6	Predictive accuracy of the user-friendly Algorithm 2 for OHS_600	285
6.7	Predictive accuracy of the take-the-best Algorithm 3 for OHS_600	286
6.8	Predictive accuracy of the maximally efficient Algorithm 4 for OHS_600	287
6.9	Predictive accuracy of the maximally efficient Algorithm 4 for OHS_extra without grammatical words	289

6.10	Predictive accuracy of the maximally efficient Algorithm 4 for Master_5+_tendency without grammatical words	289
6.11	Predictive accuracy of the maximally efficient Algorithm 4 for CompText without grammatical words	289
6.12	The maximally efficient Algorithm 4a for all parts of speech	291
6.13	Example test item for the native speaker study	292
6.14	Proportion of correct predictions by the CompSpell algorithm and two English native speakers for OHS_600 and 100-word samples from OHS_extra and Master_5+_tendency	293
6.15	Native speaker spellings differing from the unanimous dictionary spellings for OHS_600	295
6.16	Exception principles to the maximally efficient spelling algorithm	300
6.17	Simplified exception principles to the maximally efficient spelling algorithm	302
6.18	Predictive value of Algorithm 4b (including stress) for OHS_600	303
7.1	The graphical realisation of punctuation indicators (following Huddleston and Pullum 2002: 1725–1726)	308
7.2	Sepp's (2006: 86) seven logical distribution patterns for English compound spelling	311
7.3	Selected OHS_600 and Master_5+ compounds with spelling-sensitive frequencies across dictionaries	315
7.4	Number of OHS_600 compound spellings predicted correctly by rule-based algorithms and analogical variables	327
8.1	The CompSpell algorithm	352
A.1	The 200 OHS_600 compounds with exclusively open spelling	356
A.2	The 200 OHS_600 compounds with exclusively hyphenated spelling	357
A.3	The 200 OHS_600 compounds with exclusively solid spelling	359
A.4	The ninety-five biconstituent CompText compounds with open spelling	361
A.5	The twenty-one biconstituent CompText compounds with hyphenated spelling	361
A.6	The thirty-three biconstituent CompText compounds with solid spelling	362

A.7	English lexical suffixes and final combining forms (adapted from Sanchez 2008: 137–139)	362
A.8	English prefixes and initial combining forms (adapted from Sanchez 2008: 136)	364
A.9	Significant variables for OHS_600 compound spelling	365
A.10	Spelling tendencies without statistical backing in OHS_600	371
A.11	Grammatical compounds and their spelling in the <i>Oxford English Dictionary</i>	371

Acknowledgements

However large the amount of work one puts in oneself, a study such as this one is impossible without a considerable amount of support, and I would like to express my gratitude to everyone who contributed to its realisation in some way or another.

This research was carried out as a Habilitation project at the University of Erlangen-Nuremberg. I am very grateful to my supervising committee (Thomas Herbst, Mechthild Habermann and Christoph Schubert) and to my external reviewers for their helpful comments. I am particularly indebted to my academic teacher Thomas Herbst, whose insightful critical suggestions greatly improved this work. Needless to say that all remaining flaws and inadequacies are my own.

For the empirical study, the following publishers and institutions kindly answered my questions about English compound spelling and generously allowed me to use electronic lemma lists from their dictionaries, for which I am very grateful:

- Cambridge University Press: *Cambridge Advanced Learner's Dictionary* (2008)
- HarperCollins: *Collins English Dictionary* (2004)
- Langenscheidt: *Taschenwörterbuch Englisch–Deutsch* (2007)
- Langenscheidt/Collins: *Großwörterbuch Englisch–Deutsch* (2008)
- Lexicography Masterclass: *New English–Irish Dictionary* (www.focloir.ie)
- Longman/Pearson: *Longman Dictionary of Contemporary English* (2009)
- Macmillan: *Macmillan English Dictionary for Advanced Learners* (2007)
- Oxford University Press: *Oxford Advanced Learner's Dictionary* (2005)

Some of the lists kindly provided by the publishers included additional material, e.g. information on word stress (Macmillan) or on which lemmas were coded as compounds in the publisher's database (Longman).

I am also very grateful to the following researchers for giving me access to their corpora and answering my questions regarding the compilation of these resources:

- Paul Baker and Andrew Hardie: *BE06*
- Geoffrey Leech, Paul Rayson and Nicholas Smith: *BLOB-1931* (untagged)
- Craig Martell: *NPS Chat Corpus* (text version)
- Caroline Tagg: *CorTxt Corpus*

I would like to express my gratitude to Andrew Hardie for granting me access to Lancaster University's BNCweb and for his help with several technical matters regarding the extraction of information from the Lancaster corpora. My particular thanks go to Sebastian Hoffmann, who wrote a Perl script that extracted several types of information from the British National Corpus. Thomas Proisl kindly supplied a list of irregular English word forms and provided support with lemmatisation. Many thanks are also due to Peter Uhrig, who supported this research in manifold ways, from technical support (e.g. in the acquisition of data) and advice (e.g. on mark-up codes) to valuable discussions on English compound spelling.

I would like to thank Lennart Schreiber (www.webzap.eu) for writing the program CompSpell for the linguistic analysis of the data. I am very grateful to the University of Augsburg and the University of Erlangen-Nuremberg for their generous financial support of the programming work.

The statistical analysis of the data profited immensely from the discussion of statistical matters with Antony Unwin, who accompanied the project with his expertise, answered my numerous questions and contributed several graphics and statistical test results to this study. Furthermore, I would like to thank Regina Staudenmaier and Eva Fritzsche for their support with SPSS.

I would also like to express my gratitude to Ekaterina Ilieva, Stoyan Ivanov and Jelena Radosavljevic for their technical support in the preparation of some of the figures.

Several native speakers of English were so kind as to answer my questions on the acceptability of various constructed sentences and the possible modifications of potential compounds. In this context, I would like to thank the staff in the department EngPhil at the language centre of the University of Erlangen-Nuremberg, particularly Jonathan Beard and Jennie Meister. Furthermore, I would like to thank two anonymous native

speakers of British English who served as a control for the automated spelling algorithm.

I would like to express my particular gratitude to Merja Kytö, Hans-Jörg Schmid and Anatol Stefanowitsch for their detailed feedback on this volume, and to Helen Barton, Abigail Neale and Stephanie Taylor at Cambridge University Press for their kind and patient support in the final stages of publication. Thank you very much as well to Mathivathini Mareesan and Ami Naramor for their meticulous copyediting. In the course of my research project, I discussed aspects of English compound spelling with many linguists and other researchers, who recommended literature, sent me pre-published manuscripts or provided critical comments, challenging arguments or stimulating suggestions. Among those to be mentioned are (in alphabetical order) Wolfram Bublitz, Ernst Burgschmidt, Claudia Claridge, David Crystal, Stefan Evert, Ingrid Fandrych, Hans Rainer Fickenschner, Dieter Götz, Ulrike Gut, Roland Hartmann, Alasdair Heron, Dieter Kastovsky, Thomas Kohnen, Victor Kuperman, Gunter Lorenz, Christian Mair, Ingo Plag, Ulrike Stadler-Altmann, Ingrid Tieken-Boon van Ostade, Wolfgang Walther and Wolfgang Worsch. I am also grateful to the British Cabinet Office for answering my questions on spelling norms followed by the British government.

Last but not least, I would like to thank my family for their endless support. Thank you to my parents, Inge and Eduardo, also and to my parents-in-law, Margot and Peter, for everything they did to support this project over a period of several years. I am particularly grateful to my husband, Philipp, for many inspiring discussions and for his continued support. Special thanks go to our children, Sophie and Julius, for their patience and understanding as well as for making the time of writing this book very special for me.

Introduction

In view of the large amount of lexicological research in present-day English linguistics, surprisingly few studies are devoted to the spelling of English compounds (cf. 1.1.1). This is particularly striking if one considers the pervasiveness of the phenomenon, since any compound in any text necessarily requires the selection of one spelling variant (usually either as an uninterrupted chain of letters or with an intervening hyphen or space). Most of the time, language users seem to select the spelling automatically without the decision process reaching the level of conscious awareness, and conscious reflection is frequently inconclusive in those cases to which attention is drawn. The spelling of English compounds is a strikingly common difficulty and concerns everybody who produces texts in English: native speakers as well as learners of the language. As a consequence, a number of general ideas about the spelling of English compounds exist and can commonly be found in the literature. Most of these ideas can be subsumed under three general statements:

1. The spelling of English compounds follows no rules.

This view is certainly the most widespread one, and it is frequently expressed as a complaint: since no principles seem to underlie the spelling of English compounds, they are considered an important source of error and confusion. As a consequence, language users are frequently advised to consult an up-to-date high-quality dictionary when doubting about particular compounds' spelling. The following quotations are representative of numerous similar passages:

- "Of all the questions which arise in an editorial office, one of the most common has to do with compound words. Should it be written taxpayer, tax-payer, or tax payer? Solid, hyphenated, or open?" (Strumpf and Douglas 1988: 52).
- "There often seems to be little logic or consistency in the hyphenation of words in English. The reason for this is that there are few 'rules' as

such – hyphenation is often more a question of style and common sense, rather than principle” (Cullen 1999: 51).

- “Because of the variety of standard practice, the choice among these styles for a given compound represents one of the most common and vexing of all style issues that writers encounter” (Merriam-Webster 2001: 99).
- “Nowhere (or should that be ‘no where’?) is English more chaotic than in its seemingly arbitrary spelling of compound words and phrases” (Wilbers 1997).
- “The chaos prevailing among writers or printers or both regarding the use of hyphens is discreditable to English education” (Fowler 1926: 243). “[U]sage is so variable as to be better named caprice” (Fowler 1921: 7).
- “One of the most common spelling issues involve [*sic*] compound words . . . While consulting a dictionary is usually the best answer, some of these words are not necessarily found in dictionaries, and in a few cases, even if they are the spellings vary. There is also no definitive collection of rules regarding the hyphenation of compound words” (Reiser 2007).

All these statements suggest that there are no general principles offering support in the spelling of English compounds. However, it seems that most language users write the majority of compounds without previously checking their spelling in reference works. Advanced spellers in particular tend to have fairly strong intuitions about how to spell – and sometimes how not to spell – certain compounds. Since this intuition must have some kind of basis, the present study sets out to determine what makes spellers choose one variant over another. The underlying assumption here is that compound spelling is governed by certain principles, even though they can be expected to be numerous and far from obvious – otherwise they would have been discovered and made generally known long ago. At the same time, the spelling of compounds is not entirely fixed. Bauer (2003: 134) illustrates the randomness of English compound spelling by listing the spelling variants of *girlfriend* in different dictionaries: *girlfriend* in *Hamlyn’s Encyclopedic World Dictionary*, *girl-friend* in *The Concise Oxford Dictionary* (7th edition) and *girl friend* in *Webster’s Third New International Dictionary*. In addition, the most common way of spelling a particular compound may change over time – which provides a direct link to the second general statement:

2. The spelling of individual English compounds usually develops from open via hyphenated to solid spelling.

The second most important general idea about the spelling of English compounds is that they usually start off their life with the constituents separated by a space (*girl friend*), then go through a hyphenated stage (*girl-friend*) and finally finish as a solid, uninterrupted sequence of letters (*girlfriend*). Therefore any theory that attempts to discover principles underlying the spelling of English compounds needs to accommodate diachronic developments. Again, various references to the idea can be found in the literature, particularly in grammars and style guides:

- “Most compounds graduate, so to speak, from separation, through hyphenation, to integration; and everyone is entitled to his own opinion on the present status of any one of them. Thus there is still, after centuries of use, no agreement on the correctness of, say, *common sense*, *common-sense* or *commonsense*; *good will*, *good-will*, or *goodwill*” (Carey 1957: 24).
- “With compounds, the constituents are often written as separate words when the collocation seems relatively unestablished: ... As the sequence of items becomes more established, it may be hyphenated (especially in BrE) as an intermediate stage before being written solid” (Quirk et al. 1985: 1537).
- “As we have seen, compounds start as separate words, then acquire a hyphen, and end as continuous (or unbroken or solid) words” (Partridge 1953: 138).
- “**Compounds** are sometimes said to progress from being spaced as separate words, to being hyphenated, and then set solid, but the pattern is far from universal. In American English they may skip the hyphenated stage ... ; and some, especially longer ones like *daylight-saving*, may never progress beyond the hyphenated stage (in British English, or spaced, in American), however well established they are” (Peters 2004: 119).

The last quotation ends with a statement representative of the third common view regarding English compound spelling:

3. British English uses more hyphens in the spelling of compounds than American English.

- “In American English they [= compounds] may skip the hyphenated stage” (Peters 2004: 119).

- “American authors tend to use fewer hyphens than the British do” (Butcher 1992: 154).
- “In AmE, hyphenation is less common than in BrE, and instead we find the items open or solid (more usually, the latter) where BrE may use a hyphen” (Quirk et al. 1985: 1569).

It would thus seem important to take geographical considerations into account as well. The present study focuses on compound spelling in British English and draws comparisons with American English.

To conclude, the spelling of English compounds varies from both a synchronic and a diachronic perspective. The present study attempts to determine whether this variation is item-specific or whether it follows more general principles.

1.1 The State of the Art in English Compound Spelling

The spelling of English compounds is treated in extremely heterogeneous sources. The following sections provide a brief overview of its coverage in the linguistic literature (cf. 1.1.1) and the literature consulted by language users seeking advice on the spelling of particular compounds: style guides for prescriptive rules (cf. 1.1.2), grammars for descriptive rules (cf. 1.1.4) and dictionaries for item-specific information (cf. 1.1.3). Furthermore, spellers may consult a corpus (1.1.6) or use spellcheckers when typing text in word processing programs (cf. 1.1.5).

1.1.1 *Linguistic Studies*

Linguistic accounts of English compound spelling most frequently discuss spelling as a possible criterion of compounding, e.g. in treatments of word formation such as Plag (2003) or Bauer (1983). Empirical studies of the phenomenon are rare but existent.

In the pre-computerised age, in the earliest widely known study of English compound spelling, Morton Ball (1939: 46–52) assembles compounds whose spelling has changed in six editions of Webster dictionaries. A second list (Morton Ball 1939: 54–59) compares dictionaries by different publishers and “illustrates very clearly the chaotic inconsistency of general practice” (Morton Ball 1939: 43). Morton Ball’s (1939: 43) conclusion that the development “has been illogical and in absolute conflict with current usage” in many cases is indicative of the relatively prescriptive style of the book. The author does not intend a quantitative description of compound

spelling, but aims to offer stylistic guidance. It is unclear to what extent the lists are representative of the dictionaries as a whole and how the data – which are not analysed statistically – were selected.

The most important predecessor of the present study is Sepp (2006). To avoid “the gray area of the adjective+noun combinations in English . . . and obtain a less controversial set of compounds” (Sepp 2006: 22), the author restricts herself to nominal noun+noun compounds (unusually, including pronouns and acronyms), which are automatically extracted from two American English corpora that jointly comprise about 14 million words. Solid compounds are gathered by first searching for strings based on a dictionary word list and then by searching for the remaining string. Sepp focuses on the 707 compound types which occur more than thirty-five times in her corpora (Sepp 2006: 78–86). Forty-five per cent of these vary in their orthographic form: 43 occur in all three spellings, 101 have solid or open variants, 161 are either hyphenated or open, and 13 may be solid or hyphenated. Sepp checks a variety of parameters – among them the number of syllables in the compound, compound stress and double consonants across the internal boundary – and finds no single determining feature, but a cumulative effect: ten lexical features account for 66.5 per cent of the variance in solid compounds, nine features for 40.5 per cent in hyphenated compounds and ten features for 67.7 per cent in open compounds (Sepp 2006: 105–106, 109).

Mondorf’s (2009, 2000) descriptive linguistic treatment of the phenomenon is the by-product of a study on comparison in adjectives. As a consequence, Mondorf’s research is also confined to a single part of speech. She finds that about 85 per cent of adjectival compounds are hyphenated (Mondorf 2009: 378) and agrees with Sepp (2006) that the spelling of English compounds is presumably based on a large number of interacting factors.

Plag et al. (2008) consider the spelling of English noun+noun compounds regarding the effect on stress placement. Based on Google frequencies (which treat hyphens as spaces; cf. Plag et al. 2008: 776), they find “that compounds written in one word . . . are more frequently left-stressed than compounds written in two words” (Plag et al. 2008: 778).

Rakić’s (2009: 60) study also investigates only noun+noun compounds, but his material comes from a dictionary rather than a corpus: 5,270 compounds were “excerpted from” the *Longman Dictionary of Contemporary English* (LDOCE), but by what method is not elucidated. It also remains unclear why his modern British English material from a single source – whose spelling could also represent a house style – is linked

to American English frequency data from the 1960s, more specifically the Brown Corpus (Rakić 2009: 72). Rakić mainly focuses on morphological structure and finds that morphological complexity and compound length favour open spelling.

The most recent large-scale study of English compound spelling is Kuperman and Bertram (2013). Like most of its predecessors, it is limited to noun+noun compounds, but additionally excludes compounds with a final constituent consisting of verb+ *-ing*, such as *house+warming* (Kuperman and Bertram 2013: 944). Kuperman and Bertram (2013) use two corpora: the regionally indeterminate 2008 Wikipedia corpus (with 1.2 billion tokens) and a diachronic corpus of American English containing 1.8 million documents from the *New York Times* between 1986 and 2007 (Kuperman and Bertram 2013: 943–945), which may be strongly influenced by house style. Their detailed study – which restricts itself to compounds occurring in alternative variants (Kuperman and Bertram 2013: 941) – takes a multitude of factors into account in order to “identify orthographic, phonological, semantic and distributional explanatory factors” for spelling preferences (Kuperman and Bertram 2013: 943).

The remaining empirical linguistic research related to the spelling of English compounds can generally be subsumed under the heading of psycholinguistic or cognitive studies. These investigate various aspects regarding the processing of open, hyphenated and solid spelling such as the following:

- **Is it favourable to insert spaces into normally solid English compounds?** According to research by Juhasz, Inhoff and Rayner (2005), conserving the open spelling of compounds usually spelled that way leads to fast results for first fixations on the compound and lexical decision tasks. The insertion of spaces into normally solid compounds (e.g. *softball* > *soft ball*) speeds up first fixations on the compound and lexical decision. However, when refixations on the compounds are considered, spelling a solid compound open significantly disrupts processing, particularly if it consists of adj+n (rather than n+n) – presumably, because that is frequently accompanied by a change in semantics, as in the example *soft+ball*. The results indicate that open spelling facilitates access to the constituent lexemes, whereas solid spelling benefits the specification of full compound meaning, because “the direct look-up of the whole compound” is “able to begin in parallel to accessing the first constituent of the compound” (Juhasz, Inhoff and Rayner 2005: 314). All this is in line with the more general

assumption that new compounds, whose meaning can often be deduced from their parts, tend to be spelled open, whereas solid spelling is frequently found in compounds with a more specialised, lexicalised meaning (cf. 5.12.4).

- **Is it favourable to insert hyphens into normally solid compounds?** Bertram et al. (2011: 537) find an identical effect to that described earlier for spaces: the insertion of a hyphen into usually solid compounds benefits early processes and disrupts later processes. However, Bertram et al.'s (2011) results are based on triconstituent compounds, which are longer and structurally more complex than the compounds in Juhasz et al. (2005), and based on compounds from two languages other than English (Dutch and Finnish). While some results can be transferred between languages, others – particularly those referring to differences in orthographic systems – may not be so easily applied to other languages, since the expectations of readers in a language that obligatorily concatenates all compounds will presumably differ from those of readers in a language that permits more variation in the spelling.
- **Is it favourable to remove spaces from normally open English compounds?** When Juhasz et al. (2005) also examined compounds from the reverse perspective, their subjects did not find it very disturbing if usually open compounds were spelled solid – in any case less so than if solid compounds were spelled open. Furthermore, readers' eyes land further towards the centre of solid compounds, which reduces the tendency towards refixations.
- **Is the spelling of English compounds influenced by the spelling of morpho-semantically related compounds?** De Jong et al. (2002) answer this question in the affirmative after considering both the position family size and the position family frequency (cf. 5.10.3) of English compounds. They find that a large number of solid and hyphenated family members leads to faster responses from participants. The number of family members with open spelling, by contrast, seems to be irrelevant or even inhibitory, since the words with few open family members were responded to faster than those with many.
- **Are there any differences in the processing of open, hyphenated and solid compounds?** In the same study, de Jong et al. (2002) conclude that English open compounds are not part of the morphological families of simplex words and that the space between the constituents of open compounds renders their processing more comparable to that of simplex words, since it is also influenced by position family size rather than position family frequency. This outcome suggests that open

compounds may be represented differently from solid and hyphenated compounds at a central level.

- **What is the best visual cue for compound constituent boundaries?** Research for German by Inhoff, Radach and Heller (2000: 45) suggests that spacing is a particularly good visual cue to indicate constituent boundaries.

The collective view emerging from the psycholinguistic research data seems to be that open, hyphenated and solid spellings each have advantages and disadvantages for language processing. The experimental studies can only consider a relatively limited number of research items, and there is a strong focus on noun compounds only. Furthermore, the theoretical background of the phenomenon (such as the issues of norm, variation and language change) is only considered to a limited extent. Most importantly, all the studies only investigate the spelling of English compounds from a receptive perspective and do not take production into account, although there are important differences: thus readers may accept any out of two or even more spelling variants without comprehension problems (and possibly even without awareness of potential variation), but production necessarily requires the selection of the single spelling variant deemed most suitable, so that minor differences in acceptability have larger repercussions on compound spelling compared to reading.

1.1.2 Style Guides

The majority of books dealing with the spelling of English compounds are style guides treating it as one among many other issues, e.g. Bailey's (1979) *English Punctuation in Brief*, Carey's (1958) *Mind the Stop* and McDermott's (1990) *Punctuation for Now*, all of which focus on punctuation. Listing all style guides on the market is almost impossible, since many of the large publishing companies and newspapers have their own usage guides, e.g. *Webster's New World Guide to Punctuation* (Strumpf and Douglas 1988) or *Chambers Guide to Punctuation* (Cullen 1999). Style guides (which are also called *usage guides*) provide easily accessible information on constructions¹ of unclear usage (which is more difficult to retrieve from descriptive reference grammars; cf. Busse and Schröder 2010b: 99). Since style guides are almost exclusively prescriptive, and rarely indicate

¹ The term *construction*, understood in the sense of "a pairing of form with meaning/use" (Goldberg 1996: 68), is used here to refer to linguistic entities without the necessity of specifying whether these are categorised as compounds or phrases or as situated on a gradient.

their sources, they will merely be considered as the subject of investigation here, e.g. with regard to the criteria proposed for compound spelling (cf. Chapter 5) and with the aim of empirically testing the spelling rules advocated therein.

Since most style guides are published by some authority, they often describe the in-house style of a particular publisher or editorial board – e.g. *New Hart's Rules* (Ritter 2005a) for Oxford University Press. Such style guides represent the conventions agreed upon by particular relatively small groups, and they sometimes contradict each other. While they are arbitrary to a certain extent, there seems to be a common core of agreement: some rules, such as the hyphenation of ad hoc compounds, are cited by practically all books in this category – but not by Peters (2004). Her *Cambridge Guide to English Usage* is exceptional by basing its advice on corpora and questionnaires to determine acceptability, but the style guide format forces Peters to omit her statistical basis. In addition, hyphenation is treated only relatively briefly, because the style guide covers many different areas of the English language. An important dilemma of usage guides is that they have to “cover a vast amount of linguistic ground, without necessarily being an expert in all areas” (Busse and Schröder 2010a: 49). As far as hyphenation is concerned, however, the author of the classic English style guide *A Dictionary of Modern English Usage* (1926), Henry Watson Fowler, had published an almost verbatim tract on hyphenation in 1921.

Beside these shorter treatments of English compound spelling, there are two whole books devoted to the topic by Alice Morton Ball, *Compounding in the English Language* (1939) and *The Compounding and Hyphenation of English Words* (1951). The former can be considered a digest style guide discussing the treatment of English compound spelling in an impressive number of different reference works. Morton Ball also sets up her own spelling system for the *Department of State Style Manual* (1939: 67), the predecessor of the US Government Printing Office's *Style Manual* (2008), whose chapter 6 still deals with compounding rules. In Morton Ball's own prescriptive judgement, hers is “the only complete rational system that has ever been formulated up to the present time” (Morton Ball 1939: 66). A disadvantage of her work is that some rules include so many conditions that they actually correspond to three or even more rules, e.g. when she states that a “noun or adjective and a verb, used jointly as a verb; and two verbs, or a verb and a noun, used jointly as a noun” belong in the category of “Words Properly Compounded” (Morton Ball 1951: 7).

Style guides differ in the amount of detail that they offer on compound spelling in English, regarding both the space and the number of examples

devoted to the phenomenon. Thus Peters' (2004) discussion comprises little more than one page (259–260) and is limited to compounds with two constituents, while Merriam-Webster (2001) extends the treatment of compound spelling over more than twelve full pages (98–111), differentiated by part of speech and supported by additional sections on hyphens and hyphenated compounds, drawing on citation files of 15 million examples and using hedges such as *usually*, *generally*, *sometimes*. Surprisingly, Bell's (2009) *Rules and Exceptions of English Spelling* does not discuss compound spelling at all, and the prescriptive classic *Eats, Shoots and Leaves: The Zero Tolerance Approach to Punctuation* merely treats hyphenation in a “teeny-weeny, hooked-on, after-thought-y chapter” (Truss 2003: 168), in which the author expresses her regret at disappearing (and also misplaced) hyphens, whereas the question when open and solid spelling are appropriate is not discussed at all. The American classic *The Elements of Style* also limits itself to a single piece of advice in this respect: “Do not use a hyphen between words that can better be written as one word: *water-fowl*, *waterfowl*. Common sense will aid you in the decision, but a dictionary is more reliable” (Strunk and White 2000: 35).

To sum up, style guides usually offer one or more of the following types of information:

- a) prescriptive rules or descriptive principles on how to spell compounds
- b) the spelling of a relatively small number of prototypical compounds
- c) the spelling of a relatively small number of particularly frequent compounds
- d) the spelling of a relatively small number of particularly difficult compounds
- e) exceptions to the rules/principles

whereas:

- f) the spelling of individual compounds in general

is relegated to dictionaries – even if some style guides, such as Morton Ball (1939, 1951), also offer long lists of compounds and may be considered hybrid in this respect.

1.1.3 Dictionaries

Although dictionaries cannot inform on the spelling of all compounds in a language, they are far more comprehensive than style guides in this respect, which leads Strumpf and Douglas (1988: 49) to reach the (debatable)

conclusion that “[t]he surest way to determine if a word should be hyphenated is to use the dictionary”. Since spelling is an inherent feature of any compound’s lemmatised written form, dictionaries automatically contain information on their headwords’ spelling. An unusually explicit treatment can be found in the *New Oxford Dictionary for Writers and Editors* (Ritter 2005b): information on compound spelling is provided in brackets at the end of the entries or after the headwords, e.g. “**fly half** (two words)”, “**fly-past, fly-post** (hyphen)”, “**flywheel** (one word)” (Ritter 2005b: 136–137) – but how precisely those preferred variants were determined is not mentioned anywhere in the volume. The alphabetical ordering in print dictionaries and the search function in electronic dictionaries make the spelling of specific compounds easy to retrieve, whereas the use of style guides or grammars for the same purpose requires the prior analysis of the rules or principles applying to the compound to be spelled. In contrast to other reference works, which group items, dictionaries can do full justice to spelling idiosyncrasies, as each word may have its own entry.

Compound spelling concerns lexicography at various levels, beginning with lemma selection. While solid or hyphenated compounds are obvious candidates for inclusion, this is less clear for open compounds, as the borderline to phrasal constructions (which are syntactic and thus not to be included) is fuzzy (cf. 2.1). According to Stein (1985: 38):

lexicographers usually take into account their internal semantic structure. The more the meaning of a combination is assumed to be inferable from the meaning of its constituents listed in the dictionary and the process of formation itself, the stronger the likelihood that it will not be listed as a dictionary item.

One may therefore agree with Van der Colff (1998: 199) that dictionaries face the difficult situation of having to consider debatable phenomena in language whose treatment cannot be based on a generally accepted linguistic model.

If a compound is to be included in a dictionary, the next question for the lexicographer is which spelling to choose, particularly if several variants are commonly used – which is very frequently the case, according to Ritter (2005a: 42). One solution mentioned by various lexicographers is to keep the spelling from an earlier edition of the same dictionary or another, larger dictionary, particularly if little time has passed since its publication. Lexicographers may also follow the rules of an in-house style guide, as publishers often have a preference for one specific system in cases where the spelling is not fixed (Stein 1985: 41). While reference works may thus

contradict each other (cf. the comparison in Morton Ball 1939: 28), only few dictionary users are likely to notice this, as most users presumably consult only one dictionary and unquestioningly accept the spelling they find there. Since modern dictionaries attempt to record current common usage, the spelling of the lemmas is commonly based on corpora. The spelling of the lemmas in the *New Oxford Spelling Dictionary*, for instance, was determined with “the biggest ever body of English used for the purpose”, namely “the Oxford English Corpus . . . containing many hundreds of millions of words” (Waite 1995: vi).

Dictionaries may indicate variant spellings, but not always: even the *New Oxford Spelling Dictionary* with its explicit focus on orthography does not normally show compound spelling variants and tends to list only hyphenated or solid compounds (Waite 1995: viii). As for the order of alternatives, the first spelling listed will be interpreted as the preferred or dominant one (Hanks 1988; Achtert 1985: 43), and naïve language users may assume that a dictionary’s failure to list a particular orthographic variant implies that variant’s incorrectness. This is because dictionary users commonly seem to consider reference works as prescriptive verdicts on the language, whereas most present-day English lexicographers regard themselves as recording and describing linguistic usage. There seems to be an understanding “that they will never list all three spellings – solid, hyphenated, separate words – of a compound even if actually occurring in the language” (Stein 1985: 41), probably due to the immense number of entries that would be affected by such an editorial decision.

Another potential impact of compound spelling on dictionaries concerns the listing of the entries: thus in the *Longman New Universal Dictionary* (Procter 1982: xvii), a “compound written as a single word comes before the same compound written with a hyphen, which in turn comes before the same compound written as two or more separate words:

rundown *n*

run-down *adj*

run down *vt*”.

This implies that even in a printed dictionary, retrieval may be hampered by the ordering of the lemmas due to the status conferred to blanks and hyphens. While a strictly alphabetical sequence such as

hen
hen-house
hen night
hensure

disregards any inserted blanks or hyphens, the ordering in

hen
hen-house
hensure
hen night

is indicative of a system in which wordhood in the sense of an uninterrupted sequence of characters (including hyphens) overrides alphabetical ordering and may thus confuse unaware dictionary users. Yet another conceivable order is the inclusion of all compounds as run-on entries under their first element (Osselton 2005: 161). While dictionaries in the past (e.g. Johnson's 1755 *A Dictionary of the English Language*) presented hyphenated words inconsistently (e.g. as normal entries, in separate alphabetical clusters or in a series of quotations arranged alphabetically within the lemma), present-day dictionaries tend to use strict alphabetical order (Osselton 2005: 161–163). This is the most user-friendly variant, because the item in question is listed in the same place regardless of the spelling variant used.

Even large dictionaries do not list all the compounds of a language (Merriam-Webster 2001: 99) – an important disadvantage for those seeking advice on the spelling of recent or less frequent compounds. The front matter and additional sections in learner's dictionaries sometimes provide generalising information: thus OALD 7's appendix R61 outlines the principle that fractions used with uncountable or singular nouns generally combine with a singular verb (e.g. ***Fifty per cent of the land is cultivated***). It is conceivable to include general information on the spelling of English compounds in such additional sections, but this is not the kind of information one would expect in a dictionary (so that users will presumably not look for it there) but rather in a grammar.

1.1.4 Grammars

Grammars attempt to describe the systematic patterns underlying the use of a particular language, its *grammar*. The scope of the term varies in linguistics: according to Görlach (1999: 99), *grammar* is traditionally

divided into “orthography,² etymology, syntax and prosody”, but Quirk et al. (1985: 12) restrict it to the domains of syntax and inflection and point out that its meaning can be extended to comprise spelling and lexicology.

Compound spelling is an issue that touches upon different areas of linguistics without being at the core of any one of them: it is part of word formation (which centres on processes of creating new lexical items), it concerns spelling (which prototypically deals with sound–letter correspondences) and it also belongs to punctuation (whose focus lies on the level of phrases and sentences rather than words). That may be the reason why Quirk et al. (1985) merely include this issue in the appendices of their grammar: Appendix I, *Word-formation*, contains a section on “Spelling and hyphenation” (Quirk et al. 1985: 1536–1538), and Appendix III, *Punctuation*, another one on “The hyphen” (Quirk et al. 1985: 1613–1614). The chapter headings in Ritter (2005a) also differentiate between “Spelling and hyphenation” and “Punctuation”, whereas Hanks (1988) does not consider the use of the hyphen as part of spelling. Huddleston and Pullum (2002: 1760–1762) treat the spelling of compounds in their chapter on punctuation, whereas the *Advanced Grammar in Use* (Hewings 2005) does not discuss it at all. Hyphenation often assumes a more prominent role in the discussion of compound spelling than the other two variants, and the phenomenon as a whole is sometimes treated under the mere heading of *hyphenation* (e.g. in Swan 2005: 559–560). Quirk et al. (1985: 1613–1614) list nine compound types which are commonly hyphenated (e.g. noun compounds with an adverbial second constituent, such as *runner-up*), but they do not provide anything similar for open or solid compounds.

In contrast to dictionaries, grammars have the potential to abstract and consequently to provide principles applying to many linguistic items, which represents an economical advantage compared with the one-by-one lexical approach in dictionaries. As regards English compound spelling, grammars tend to stress the difficulty of defining the compound concept, to give selected examples and to point out that the spelling is not always regular. In view of its conclusion that “there are no rules” for English compound spelling, the *Longman Advanced Learners’ Grammar*

² In this context, the term *orthography* is understood as the “part of grammar which treats of the nature and values of letters and of their combination to express sounds and words; the subject of spelling” (OED s.v. *orthography* 1. b.). In the remainder of the present study, by contrast, the term *spelling* will be used to refer to such descriptive aspects, and *orthography* will be used in its more common prescriptive meaning: “[c]orrect or proper spelling, spelling according to accepted usage; the way in which words are conventionally written” (OED s.v. *orthography* 1. a.).

advises its readers that “it is best to check in an up-to-date dictionary” (Foley and Hall 2003: 259). This is representative of the current situation, in which the apparent lack of generalizable principles has led to the common perception of English compound spelling as a peripheral aspect of English grammar – if part of grammar at all.

1.1.5 Spellcheckers

Since a large proportion of present-day written communication is computer-mediated or originates on a keyboard and screen, written text production is frequently assisted by spellcheckers, e.g. in word processors. Spellcheckers differ from the reference works discussed earlier in this chapter in that they are frequently used automatically (not necessary intentionally) and do not provide a surface form of the compounds until these are typed in by the users themselves.

Spellcheckers retrieve missing or superfluous characters, scrambled letters etc. even more accurately than most humans, but only as long as these can be retrieved based on the principles that the programs follow. Words – which are defined as sequences of characters between two blanks or a blank and a punctuation mark (e.g. a full stop) in this context – are checked against an internal word list, to which users may add their own items. If a word is not in the list, it is underlined as faulty, but mistakes which result in possible words cannot be retrieved this way, e.g. *week* instead of *weak*, or *sing* instead of *sting*. With regard to compound spelling, this would mean that compounds with open spelling can never be evaluated as incorrect, because the program will usually find the two constituents in its checklist. The performance of some spellcheckers such as that in Microsoft *Word* greatly benefits from an additional grammatical component, which recognises e.g. that **I have two car* is not correct, because a noun following the numeral *two* should be pluralised, and the program will consequently underline *car* as incorrect. Such a spellchecker might recognise that the sentence *?Smith is an also ran* is unusual, because the indefinite article *an* is not followed by a noun, as expected, but by an adverb, and might mark the sentence as ungrammatical. However, this is not the case in Microsoft *Word*, in spite of the fact that the compound *also-ran* (‘someone who fails to win a competition, election etc.’; cf. LDOCE) is hyphenated in all dictionaries considered here. This suggests that open spelling is not subject to grammatical testing in *Word*, because there are so many open compounds that they cannot all be captured by a word list (particularly since it is commonly assumed that new compounds generally use open spelling). The situation

predicted by Greenbaum (1986: 263), that possibly “the growing use of wordprocessors and the accompanying spelling checkers will eventually impose complete uniformity of spelling in the language”, is thus still far from being realised, even if one may expect the most successful spellcheckers’ spelling algorithms to have some influence on English orthography.

1.1.6 *Corpora*

The treatment of compound spelling in corpora differs from that in the other resources in that corpora provide raw rather than edited material. Corpora may be used by linguists or non-linguists to check the frequency of individual spelling variants for particular compounds. Since end-of-line hyphens are formally identical with compound-internal line-final hyphens, compound spelling is an issue in corpus compilation. There are various approaches, such as keeping or deleting end-of-line hyphens generally or in accordance with some reference work (cf. 4.2).

Compound spelling also plays a role from the perspective of information retrieval in corpus research, as divergent spelling may hide some of the compounds in a corpus (e.g. when searching for *girl-friend* in a corpus containing only *girlfriend*). This can be avoided by using a wildcard standing for zero or one character in the middle of all compounds or by carrying out three individual searches to cover all variants, but not all corpus-based linguistic studies without a focus on spelling may have considered this.

These issues can be extended to digital texts, particularly those in pdf format, which are increasingly important in academic research and can be considered a kind of corpus as well. Hyphens at the end of lines in pdfs are automatically ignored by Adobe Acrobat Reader’s search function and automatically deleted if text from a pdf is copied into a word processor. This is a relatively good strategy, since most hyphens in that position are presumably end-of-line hyphens which would otherwise mask search sequences that are inserted in standard format (i.e. without a hyphen, e.g. *never* rather than *ne-ver*). However, there is the inherent danger that the search for a sequence with a hyphenated compound may incorrectly return no hits, because the compound’s hyphen occurs at the end of a line – an aspect that advanced pdf search functions of the future should ideally take into account. For instance, only the search for the unusual spelling “higherfrequency” finds the target sequence “higher-frequency compounds were more likely to be spelled as one word” in the pdf of Kuperman and Bertram (2013: 946).

1.2 Summary and Aims of the Present Study

Innumerable English compounds are written every day, and in each case, the speller needs to select a single spelling variant. The spelling of English compounds is therefore a relevant issue that many different types of texts or resources are concerned with: previous linguistic studies have examined various aspects of English compound spelling, often from a cognitive perspective. Furthermore, language users frequently draw information supporting them in variant selection from sources which do not focus on compound spelling, such as dictionaries. Style guides and grammars usually cover English compound spelling among one of many more areas. Spellcheckers which are integrated into word processors are an increasingly important type of reference, and corpora provide raw material for the frequency analysis of spelling variants.

This heterogeneity makes it necessary to outline the aims of the present account of English compound spelling and to state clearly what it is not: this is a linguistic study and no style guide. At best, the heuristics derived from the empirical study may “provide advice through **descriptive** information on usage” (Peters 2004: 150). The present account of English compound spelling will not provide an exhaustive list of all the rules mentioned in style guides, but merely discuss some of these as hypotheses that are subjected to empirical testing. The focus of the present study lies on the empirical investigation and theoretical modelling of linguistic variation, its determinants and its development for a particular aspect of the English language, namely English compound spelling.

While common usage often employs the term *hyphenation* indiscriminately to refer both to the *hard hyphen* joining constituents of compounds (sometimes also affixes and bases) and the *soft hyphen* used in end-of-line hyphenation (cf. Huddleston and Pullum 2002: 1759; Ritter 2005a: 52), only the first of these is considered in the present study, since the two phenomena differ considerably: the hard hyphen is a lexical phenomenon which may be linked to particular lexical items (although this study attempts to discover regularities underlying its use), whereas the distribution of the soft hyphen is a relatively regular syntactic phenomenon governed by generally applicable rules based either on phonological criteria (*struc|ture*) or on morphological criteria (*struct|ure*) (Quirk et al. 1985: 1613). The two types of hyphen also differ from a formal point of view (because hard hyphens are preceded and followed by letters, whereas soft hyphens are preceded by a letter but followed by a line break) and with regard to their status: while all other punctuation marks, including hard hyphens,

are conserved in verbatim quotations (e.g. in academic texts), end-of-line hyphenation is not, because it is not considered part of the noteworthy form (cf. also Sanchez-Stockhammer 2017).

The aim of the present study is to provide a comprehensive account of compound spelling in English, starting with the delimitation of the compound concept used here (cf. Chapter 2) and a discussion of the normative background of the phenomenon (cf. Chapter 3). The empirical testing of more than sixty hypotheses on the spelling of English compounds (cf. Chapters 4–5) is subsequently condensed into several descriptive algorithms which may assist e.g. learners of English in variant selection (cf. Chapter 6). Building on prototype theory, among others, theoretical and cognitive models for present-day British English are then derived from the research and complemented by the perspective of language change (cf. Chapter 7). To sum up, this volume attempts to provide an in-depth account of compound spelling in English, of its development and the many domains into which this only seemingly superficial phenomenon radiates.

The linguistic study presented here differs from its predecessors particularly in the scope of the compounds under consideration. While previous work is usually restricted to noun+noun compounds (e.g. Sepp 2006; Kuperman and Bertram 2013) or adjective compounds (e.g. Mondorf 2009), the present research takes into account all lexical entities which can be considered compounds based on a relatively general definition (cf. 2.6). This approach is firmly rooted in the assumption that a theory on the spelling of English compounds should encompass all such entities and not only those belonging to restricted parts of speech – particularly since part of speech plays an important role in the selection of spelling variants, as we shall see.

The difference in scope also extends to the regional variants of English dealt with. Thus Morton Ball (1939, 1951) and Sepp (2006) concentrate on American English only, whereas Rakić (2009, 2010) only covers British English. The present research, by contrast, considers both British and American English, but with a focus on the former. While Morton Ball (1939) and Rakić (2009) only use data from dictionaries and Sepp (2006) extracts her compounds exclusively from corpora, the present study considers both types of material and can consequently give an account both of the variation between different reference works and of the difference between lexicographic spelling information and actual language usage.

In contrast to most research, which takes compounds with varying spellings as the starting point (e.g. Sepp 2006; Kuperman and Bertram

2013), the present study begins by analysing those compounds that can be considered firmly established in their spelling in order to determine what criteria might be most important for variant selection.

In addition, this study complements the synchronic investigation of compound spelling with a diachronic, corpus-based perspective – a characteristic shared only with Kuperman and Bertram (2013), but covering a longer period of time than its predecessor (which focuses on American English).

In conclusion, this study provides a detailed theoretical treatment of English compound spelling, a phenomenon which is relevant for everybody writing texts in English. This empirical and descriptive research aims to provide one of the most detailed and comprehensive surveys of English compound spelling up to the time of writing by going beyond noun compounds (treated e.g. in Sepp 2006; Kuperman and Bertram 2013) and adjective compounds (Mondorf 2009), by using both dictionary data (Morton Ball 1939; Rakić 2009, 2010) and corpus data (e.g. Sepp 2006), by considering both British English (Rakić 2009, 2010) and American English (e.g. Sepp 2006), by offering a diachronic perspective (Kuperman and Bertram 2013) and by considering a very large number of potentially distinctive features.

PART I

Theoretical Background

The following sections provide the theoretical background for a large-scale empirical study: the category of the compound is first delimited in its scope by comparison with adjacent categories such as syntactic phrases, simplex lexemes, derivatives, multi-word items and names. The subsequent overview of compound types from various perspectives (e.g. word-formational structure and part of speech) culminates in a detailed definition of the present study's compound concept. This is followed by an analysis of the normative background within which the phenomenon of English compound spelling is situated.

Delimitating the Compound Concept

What is referred to as *compound spelling* in the present study is treated under varying names in the literature: Morton Ball's (1951: 3) use of *compounding* to refer to spelling only results in an unfortunate terminological overlap with the word formation type. The most frequently used term seems to be *hyphenation* (e.g. in Bauer 2003: 134 and the *GPO Style Manual* 2008), which unfortunately suggests the use of a hyphen. Since hyphenation emerges as the most marked of the three spelling variants (cf. 7.1), this terminology is particularly unusual, considering that generally the unmarked constructions from sets of oppositions are used to refer to a whole dimension in a hyperonymous function (e.g. neutral *length* as against marked *shortness*; cf. Leech 1981: 113–115). Furthermore, *hyphenation* is an ambiguous term, because it frequently refers to the splitting of words at the end of lines. As a consequence, the present account uses the term *compound spelling* as a general, neutral and less ambiguous term, which deals with the principles underlying the way the constituents of English compounds are combined in writing.

In order to discuss the spelling of English compounds, it is necessary to determine first what kinds of construction are recognised as belonging in that category. In spite of the multitude of books and articles dealing with compounds, Faiß' (1981: 132) observation that there does not seem to be a generally recognised definition of what constitutes a compound still holds true more than thirty years later. This is particularly problematic in view of the large number of adjacent categories which compounds need to be distinguished from (cf. 2.1–2.4), so that most previous research has focused on the centre of the category (i.e. nominal noun+noun compounds). In the following, an attempt will therefore be made to explore the boundaries of the compound concept.

As the name indicates, compounds are compounded lexemes¹ and therefore consist of more than one constituent. Bauer (1983: 29) defines a compound as “a lexeme containing two or more potential stems that has not subsequently been subjected to a derivational process” (Bauer 1983: 29). Since “it is quite common to find compound prepositions and compounds in other minor categories” (Bauer 2003: 137), it makes sense to expand the definition of the constituents beyond lexical stems, as in Quirk et al.’s (1985: 1567) definition of the compound as “a lexical unit consisting of more than one base . . . and functioning both grammatically and semantically as a single word”. This implicit inclusion of grammatical constituents results in the following preliminary definition (cf. 2.6 for the final version):

A compound is a complex lexeme which consists of at least two constituents occurring as free lexemes each and which contains no affixation on the highest structural level.

However, this definition still leaves room for interpretation. Compounds need to be set apart both from other lexemes (most of which are regarded as the result of word formation processes) and from phrases (which are regarded as the result of syntactic processes), although there seems to be a tendency to regard these two domains as ever more gradient (e.g. Erman and Warren 2000: 53). The following sections discuss the criteria which are commonly used in the literature to distinguish compounds from other linguistic entities. Furthermore, they give an overview of the types of compound recognised in the present study based on length, word formation type, part of speech and spelling.

2.1 Compounds versus Phrases

It is the distinction between compounds and phrases that seems to represent the most important difficulty – at least considering the vast amount of literature devoted to the topic. According to Donalies (2003: 79), the comparison may be complicated by differences between the categories, with words being formed according to word formation rules and phrases being formed according to syntactic rules. However, the general

¹ The term *lexeme* is understood here in the sense of “the fundamental unit . . . of the lexicon of the language” (Matthews 1974: 22) and following Cruse (1986: 80), who defines the lexeme as “a family of lexical units” – by which he understands “the union of a single sense with a lexical form”, the latter of which is an abstraction from all possible inflected forms. In addition, the present account also uses the term to subsume grammatical words, regardless of whether they can be inflected (e.g. pronouns such as *nobody*’s) or not.

assumption that compounding has its origin in a univerbation process applied to syntactic structures in Proto-Indo-European, whose output served as the basis for analogical formations without corresponding syntactic structures (Kastovsky 2009: 328–329), would rather lend support to the currently more common view in linguistics (e.g. in valency grammar and construction grammar) that there is no clear dividing line between syntax and the lexicon (cf. e.g. Herbst and Schüller 2008: 1). From a synchronic perspective, a more important reason for the difficulty experienced in the distinction between compounds and phrases may therefore be that “the lack of inflectional morphemes in English . . . makes surface forms of English compounds and free syntactic groups identical in terms of their morphological forms” (Lieber and Štekauer 2009: 5), e.g. the English compound *blackberry* and the phrase *black berry* compared to their German equivalents *Blaubeere* and *blaue Beere*. Particularly problematic are syntactically permissible constructions with initial adjectives that are still considered compounds by many scholars (e.g. the adj+n compound *black eye* as against the adj+n phrase *black shoes*). Furthermore, English has hardly any denominal adjectives denoting material in English (such as *wooden*; cf. Giegerich 2004: 7), so that nouns are often used with an adjective-like function in noun+noun constructions whose first noun denotes the material of the second one (e.g. *steel bridge*). Similarly, while English derives adjectives from place names for countries (e.g. *Spain – Spanish*; *Italy – Italian*), this is not usually the case for the name of towns, for which gradient noun+noun constructions (e.g. *an Oxford don*) are used as well.

The most influential discussion of possible criteria for the distinction between compounds and phrases can be found in Bauer (1998). While the discussion is restricted to noun+noun combinations, many of the criteria considered there can be extended to compounds combining other parts of speech. The following sections present a critical discussion of the most commonly used criteria in the linguistic literature, in which the discussion of formal aspects is followed by the consideration of syntactic, structural and finally semantic criteria.

2.1.1 *Formal Criteria*

2.1.1.1 *Orthographic Unity*

A study which sets out to test compound spelling can obviously not base its definition of the compound on orthographic unity, but it is still necessary to discuss this criterion, because it is so frequently mentioned in the

literature. Indeed, some linguists (e.g. Morton Ball 1951: 3) exclude open compounds from their compound definition altogether. However, if orthographic unity were a necessary requirement for compound status, orthographic variants with identical phonological form and meaning but different spelling would call for very distinct categorisation: *girlfriend* and *girl-friend* would have to be classified as compounds and their variant *girl friend* as a phrase (Donalies 2003: 80). Possibly for that reason most accounts of English compound spelling merely regret the fact that English compounds cannot be defined as an uninterrupted sequence of characters, but still admit open (and also hyphenated) spellings on the grounds that some patterns producing otherwise prototypical compounds (such as V+ing + N, e.g. *nursing home*) commonly use open spelling (Schmid 2011: 122). Yet while orthographic unity is no necessary defining criterion for compounds, it is a sufficient one: constructions consisting of sequences of letters which are not interrupted by a space will generally be interpreted as a single lexeme (cf. Schmid 2011: 132) and thus as compounds if their constituents are lexemes in their own right. As soon as characters other than the hyphen enter the sequence, this is no longer necessarily the case (cf. contractions such as *don't*). With both solid and hyphenated spelling indicating word status, orthographic unity is thus a good criterion for two of the three compound spelling variant types, but not for compounds in general (a large number of which are spelled open; cf. 7.2). While orthographic unity can thus not be used as a clear defining criterion for compounds, it is, however, an exclusive criterion for phrases, as syntactic groups are neither hyphenated nor written solid (Faiß 1981: 135).

2.1.1.2 *Fore-Stress*

The most common test criterion for compound status in the literature is presumably fore-stress, which distinguishes between compounds like *'greenhouses*, with stress on the first constituent, and phrases like *,green 'houses*, with stress on the second (cf. also 5.4). Marchand (1969: 25) sees a connection between primary stress on the first constituent and the permanent lexical relation expressed in compounds, and links primary stress on the second constituent to a mere syntactic relation. While this is in line with the fact that Germanic languages – including English – usually place word stress on the first syllable, the existence of French borrowings with word-final stress (e.g. *champagne*, *magazine*) in English has established “a precedent for end-stressed nouns in the lexicon” (Giegerich 2004: 6). This might explain why fore-stress is no

unproblematic test criterion for compound status in English anymore (cf. Plag, Kunter and Lappe 2007; Bell and Plag 2012):

- Some compounds, such as *blackcurrant*, *full stop* ('period') or *hotdog*, are stressed on the second constituent by many speakers (Huddleston and Pullum 2002: 451, 1650). The same applies to inversion compounds such as *heir apparent* (Faiß 1981: 133) and to almost all copulative compounds (Schmid 2011: 145).
- Speakers may apply stress placement inconsistently to the same construction (Bauer 1983: 102–104) and dictionaries may also differ in their treatment of the phenomenon: thus *churchwarden* has initial stress in one dictionary but final stress in others (cf. Bauer 1998: 70).
- For some lexical items, such as *ice cream*, there are even generally recognised alternative stress patterns (Bauer 1983: 102–104).
- Some authors argue that there is an association between particular stress patterns and particular semantic relations: thus 'B made of A' (*stone wall*) sometimes calls forth end stress, while 'B used for A' (*pruning shears*) calls forth fore-stress, but it is unclear why one of these relations should be considered more lexical than the other (Bauer 1998: 71).
- Combinations that are 'too long' always have two stresses, e.g. *concert performance* (Marchand 1960a: 16).

Taking everything into account, it seems that word stress is most distinctive in adjective+noun compounds (e.g. *blackbird* vs. *black bird*) but that fore-stress is no criterion which can distinguish all compounds from phrases. In a reversal of perspectives, however, fore-stress can be used to distinguish phrases from compounds, since phrases never have fore-stress (cf. also Giegerich 2004: 21 and Faiß 1981: 133).

2.1.1.3 Length

Another possible formal test criterion is length. Bauer (2003: 134) states that some linguists "seem happy enough to concede *GIRLFRIEND* (however spelt) as a single lexeme but are less happy with longer compounds", e.g. *morphology textbook* vs. *morphology textbook cover* and *morphology textbook cover box*. Although processing of long compounds should be easier in written than in spoken language, Schmid (2011: 208) finds only a small number of compounds with four or more constituents in his written corpus. A statement such as "*and then we looked it up in the airline cabin crew safety training manual*" is uncommon even in technical language, since

extremely long combinations with potential compound status are likely to be replaced by abbreviations or acronyms (Schmid 2011: 206). Phrases, by contrast, can be very long, e.g. due to multiple embedding. Length is thus no absolute but rather a gradient criterion.

2.1.2 Syntactic Criteria

2.1.2.1 Part-of-Speech Specification

Words and thus also compounds can be assigned a part of speech based on their inflection and the context in which they occur (Plag 2003: 8). Since compounds do not cross phrasal boundaries, they either constitute a whole phrase on their own (e.g. a verbal compound in the simple present or past tense, such as the verb phrase in *They **double-checked** the calculations*) or they occur within a phrase (e.g. the noun *fairy tale* acting as the head of the noun phrase *One could call it **a modern fairy tale***). The mere occurrence in a particular syntactic slot is not enough to classify a construction as a compound. While some constructions may be considered adjective compounds in the premodifying slot of a noun phrase (e.g. *full+length* in *a full-length portrait*), this is not automatically the case: *last year's* in *There were fifty nominees for last year's prize* is considered phrasal in view of the fact that the genitive inflection cannot be used with adjectives. However, without the help of other criteria, the line is sometimes difficult to draw: for the sentence *Next to me sat a smiling child*, few people would argue that *smiling child* is a compound. However, if somebody wished to do so, they could consider the construction a noun based on the part of speech of the head *child*, and the fact that inflection can be added (e.g. *smiling children*, *smiling child's*, *smiling children's*) would seem to support such an analysis. Part-of-speech specification is presumably most useful in the classification of constructions consisting of grammatical words, e.g. *in+as+much+as*, whose classification in the LDOCE example sentence *Ann was guilty, inasmuch as she knew what the others were planning* as a complex conjunction is supported by its possible replacement with the conjunction *because*. As a consequence, part-of-speech assignment is best used in combination with other criteria to distinguish compounds from phrases.

2.1.2.2 Syntactic Ill-Formedness

Another possible compound criterion is syntactic ill-formedness: some constructions consisting of free constituents on the highest level of analysis, such as *forward+looking* or *with+out*, cannot be described by means of syntactic rules, as one would rather expect the order *looking+forward* or

with followed by a noun phrase. As a consequence, they can be considered compounds (cf. Dressler 2005: 28). The opposite, syntactic well-formedness, by contrast, has no such implications, since the order of the constituents in compounds may also correspond to syntactically well-formed phrases (e.g. in *greenhouse* ‘a glass building for plants’). Syntactic ill-formedness is thus a possible but not necessary criterion for compound status.

2.1.2.3 Positional Mobility

Positional mobility as yet another criterion of wordhood could also be considered as potentially distinguishing between compounds and phrases: for instance, both *police officers* and *lollipop ladies* “can be used in different places in the sentence” (Bauer 1983: 105), e.g. as subjects or objects in *Police officers stopped lollipop ladies* vs. *Lollipop ladies stopped police officers*. However, in such cases it is actually whole phrases (including determiners if the noun is not uncountable or pluralised) which change their position. Since some syntactic phrases, particularly with adverbial function, are also relatively mobile within the sentence (e.g. *For that reason, he did not come* vs. *He did not come for that reason*), positional mobility is no valid criterion to distinguish compounds from phrases.

2.1.2.4 Uninterruptability

While it is possible to interrupt phrases (by extending the noun phrase *a girl* to *a nice girl* and even *a nice young girl*, or the verb phrase *was going* to *was happily going*),² “items cannot be inserted between formatives within a word” (Bauer 1983: 105), so that a compound such as the noun *girl+friend* cannot be extended to *girl+nice+friend*. At first sight, this statement seems to be contradicted by some examples that Bauer (1983: 106) gives of “complex words of the form AB such that there is also a complex word of the form ACB”, in which “the element C forms a unit either with A or with B”. However, *library book*’s supposed extension to *library comic-book* and *city office*’s supposed extension to *city insurance office* have to be rejected as instances of interruptability, because these sequences are better analysed as new compounds with partly new constituents (i.e. *comic-book* and *insurance office* as new complex constituents). Whether a construction can be interrupted thus depends on whether

² Note that it is not possible to insert random constituents in random positions within phrases either (e.g. **a nice the girl* would be ungrammatical). While this would seem to contradict interruptability’s requirement that it should be possible to insert elements “more or less freely” (Bauer 1983: 106; Lyons 1968: 204), interruptability is at least generally possible in phrases as opposed to compounds.

Table 2.1 *Order of English adjectives within the noun phrase*

evaluation	size	condition	age	shape	colour	origin	material	function or classification	NOUN
<i>nice</i>	<i>small</i>	<i>dusty</i>	<i>modern</i>	<i>round</i>	<i>black</i>	<i>German</i>	<i>woollen</i>	<i>political</i>	<i>things</i>

it represents a unified concept. While uninterruptability can be accepted as a valid criterion to distinguish phrases from compounds, there is an exception in the form of conjoints (cf. 2.1.2.7), in which the shared constituent in “two linked units of equal status” (Quirk et al. 1985: 46) is only physically expressed once, e.g. when the compound *iron bars* is superficially interrupted by *and* and *steel* in the construction *iron and steel bars*.

In practice, however, uninterruptability is not always easy to determine, particularly in syntactically well-formed combinations of adjective and noun whose status as a technical term with unified meaning (cf. 2.1.4.3) depends on previous lexical knowledge (e.g. of *modern man* as an anatomic term). While it should be possible to insert an adjective if such constructions are phrases, the test is limited by the tendency of English adjectives to occur in a particular order. Table 2.1 combines and slightly modifies the descriptions in Endley (2010: 96–97), DeCapua (2008: 94–95) and Swan (2005: 11):

Testing requires an adjective which can at least theoretically be inserted between a potential compound’s adjective and noun constituents, and which should therefore come from a category to the right of the tested construction’s adjective in Table 2.1. For instance, the status of the construction *modern man* cannot be tested by the insertion of *nice*, since the sequence [?]*modern nice man* violates the ordering constraint, which permits no conclusions concerning interruptability – in contrast to the insertion of *German*, which results in a semantic change in *modern German man*.

2.1.2.5 *Syntactic Isolation of Constituents*

According to Bauer (1998: 72–74), “elements within the word should not be available to the syntax”. For instance, the status of a construction as an adj+n compound can be tested by attempting comparison if that is permitted by the adjective on its own. As a consequence, *special education* can be classified as a compound, because the sequence *more special education* would result in a change in meaning. Yet the principle of syntactic isolation

is broken in attested examples of compounding such as *So, I hear you're a real cat-lover. How many do you have now?*, where *how many* refers to *cat*, the modifier of the compound (Bauer 1998: 72). The principle is also broken in the case of some derivatives, whose status as words is uncontested, e.g. in *what sharply distinguishes Chomskyan practice from that of his structuralist forbears*, where *his* refers anaphorically to *Chomsky*, the base of the suffixation (Bauer 1998: 72). Furthermore, some constructions which are structured in parallel to (other) compounds comprise internal inflection, e.g. *games mistress* with an initial plural (Bauer 1998: 72–73), or genitive compounds such as *bull's-eye* (cf. 5.5.1). While all this seems to imply that the syntactic isolation of constituents is no valid criterion to distinguish compounds from phrases, one may argue that the inflection is actually part of these latter compounds and can therefore not be modified or deleted to form a commonly used singular or non-possessive form (*?game mistress*, *?bull-eye*). Another group of seemingly contradictory compounds is represented by constructions like *sons-in-law*, *mothers-to-be* and *lookers-on*, which are usually pluralised in the middle of a hyphenated sequence of letters (Swan 2005: 517). This group can be included by modifying the principle of syntactic isolation of constituents in such a way that compounds are expected to take inflection only at their head (Donalies 2003). However, modern-day usage also seems to permit word-final placement of the genitive for such items (*son-in-law's*, *mother-to-be's*, *looker-on's*; cf. e.g. www.englishforums.com/English/InflectedPeriphrasticGenitive/bvpjk/post.htm, 18 August 2017). Since neither word-final inflection nor head-only inflection seems to apply consistently, the central principle which can be derived from all of the foregoing discussion is that compounds only take each type of inflection once for all their constituents. However, the principle needs to be refined even further if one considers that compounds with internal inflection may occur in syntactic contexts requiring the same type of inflection at the end of the compound, e.g. *news bulletin* in the plural context *Robin enjoys watching news bulletins*. In order to avoid having to evaluate a construction's status as a compound differently depending on whether the context requires the same type of inflection as at the end of the first constituent or another type of inflection or none, the most precise formulation of the principle of syntactic isolation of constituents is that each type of inflection may only be applied once to the base form of a compound (which may contain inflection at the end of the first constituent). The only exception to this principle consists in phrases of the type 'name + *and/or* + name' and their variants with more constituents linked by commas, which take a single inflection (like

compounds) in spite of the fact that two inflections are theoretically possible. For instance, the book *Cohesion in English* is usually referred to as *Halliday and Hasan's book* and not as *'Halliday's and Hasan's book*, and the form *Jack or Jill's* seems to be more common than *Jack's or Jill's* (Google search, 04 September 2015).

2.1.2.6 *No Replacement of the Head by the Pro-form One*

While the syntactic isolation of compounds' constituents theoretically also covers the impossibility of replacing a compound's head by the pro-form *one*, e.g. in **That's not an oak tree but an elm one*, it is listed separately here because of its frequent discussion in the literature (e.g. Bauer 1998: 76–78). While Schmid (2011: 132) uses this criterion to distinguish phrases from compounds, Bauer (1998) provides a list of counterexamples, e.g. *I told you to bring me a **steel bar** but you have brought me an **iron one*** or *I wanted a **sewing machine**, but he bought a **knitting one***, which renders the distinctiveness of this criterion for the discrimination of compounds and phrases slightly doubtful.

2.1.2.7 *No Coordination*

Another traditional criterion, similar to syntactic isolation, posits that no coordination should be possible with the constituents of a compound (e.g. *butter+cup*), so that neither **bread and buttercups* nor **buttercup and saucer* are possible combinations (Bauer 1998: 74). However, the impossibility to coordinate the flower *buttercup* may be the result of the compound's idiomaticity: since coordination requires a parallel semantic relationship between the coordinated elements (e.g. like that between *buttercup* and the hypothetical flower *honeycup* in the presumably possible combination *butter(-) and honeycups*), finding possible items to coordinate becomes increasingly difficult with increasing idiomaticity (Bauer 1998: 74–75). Where such a parallel structure exists, however – and that seems to be the case for most constructions considered compounds – coordination is possible at least in noun+noun compounds (cf. Bauer's 1998: 74–75 *iron and steel bars* and *steel bars and weights*), but possibly also in verbal compounds (*deep freeze and fry*) and adjectival compounds (*high maintenance and performance*). The non-coordination of constituents is thus no absolute criterion for the distinction between compounds and phrases, either.

2.1.3 Structural Criteria

2.1.3.1 Internal Stability

According to the structural criterion of internal stability, lexical constituents “cannot be reordered within the word” without resulting in impossible variants of compounds (cf. the scrambled versions of *forget-me-not*: ²*not-me-forget* and ²*not-forget-me*) or existing but distinct words with a change in meaning (when *garden city* is reordered to *city garden*; cf. Bauer 1983: 105–107). The constituents of copulative compounds (cf. 2.5.2) are sometimes claimed to represent an exception, but it seems that these are rarely reordered in actual usage: thus *singer-songwriter* occurs twenty-four times in the British National Corpus, compared to zero hits for **songwriter-singer*. While one may therefore generalise that compounds are internally stable, the consideration of phrase structure leads to the same result: a phrase such as *the nice young girl* cannot be randomly reordered. The result would be quite unusual in some cases (²*the young nice girl*) and ungrammatical in the majority of instances (²*nice the girl young*/²*girl young the nice* etc.). It is therefore possible to conclude that internal stability is a quality of both compounds and phrases and therefore no distinctive criterion.

2.1.3.2 Right-Headedness

Donalies (2003: 84–85) and Adams (2001: 3) consider right-headedness as a potential indicator of compoundhood, but constructions of the type *mother-in-law* with an initial head lead Donalies (2003) to discard this criterion. Furthermore, neither phrase compounds nor copulative compounds (cf. 2.5.2), which are considered compounds by several treatments of word formation (e.g. Bauer 1983; Dressler 2005), fulfil this criterion. Left-headedness is observable in some phrases allowing postmodification (e.g. the noun phrase *girls united*), but since noun and adjective phrases are usually right-headed (e.g. *a strong **desire**; extremely **happy***), headedness is no useful distinction between compounds and phrases.

2.1.3.3 Listedness

The listedness of compounds is one of the most important defining criteria mentioned in the literature and may be interpreted either in a lexicographical way (e.g. by Bauer 1998: 67) or with regard to storage in the mental lexicon (e.g. by Schmid 2011: 122). The criticism which can be applied is similar in both cases: if listedness were the only criterion for compoundhood, the addition of newly formed (and therefore at least

initially unlisted) compounds after the determination of a status quo would be impossible. Conversely, both dictionaries and the mental lexicon may also list longer entities such as idioms and whole sentences (e.g. proverbs). Since rule-produced entities such as syntactic units may be listed (Langacker 1987: 29), listedness “cannot be used to set off compounds from anything else” (Bauer 1998: 68) and is therefore discarded as a criterion here.

2.1.4 Semantic Criteria

2.1.4.1 Idiomaticity

Many linguists use idiomaticity (as one aspect of listedness) to determine compound status (cf. Bauer 1998: 67): thus Kruisinga (1932: 1581) defines a compound as “a combination of two words forming a unit which is not identical with the combined forms or meanings of its elements”, and, according to Marchand (1960a: 18), compounds “denote an intimate, permanent relationship between the two significates to the extent that the compound is no longer to be understood as the sum of the constituent elements” (e.g. *butterfly*, which is clearly idiomatic). However, many compounds can actually be interpreted in a literal sense, e.g. *passenger seat* (‘a seat for passengers’) or *oven-ready* (‘ready for the oven’). Furthermore, “[a]ny syntactic group may have a meaning that is not the mere additive result of the constituents” (Marchand 1960a: 80). Thus both the Old English compound *hēafod-gim* ‘head-gem, eye’ and the parallel syntactic construction *hēafdes gim* ‘head’s gem, eye’ with an inflected first constituent have idiomatic meaning (Terasawa 1994: 73). Taking everything into account, idiomaticity cannot clearly distinguish compounds and phrases from each other.

2.1.4.2 Semantic Specificity

A related criterion is semantic specificity. According to Faiß (1981: 134), “[m]any scholars hold that a compound is semantically more restricted or more specified than a parallel syntactic group is”. Thus a *revolving door* is not simply a door that revolves but a particular kind of door (Faiß 1978: 25), and the phrase *a dancing girl* differs from the compound *dancing-girl* by the latter’s professional status (Hansen et al. 1990: 50). However, it is difficult to “formalise this intuitive distinction” in a more general way (Bauer 1978: 43) – particularly since only a small number of compounds contrast with a parallel syntactic group. As a consequence, the criterion of semantic specificity has only limited applicability for the present study.

2.1.4.3 *Unified Semantic Concept*

The idea that compounds refer to a unified semantic concept is very common in the literature (e.g. Plag 2003: 7) and seems to be generally accepted. For Schmid (2011: 142), the most important cognitive function of compounding is that compounds establish links between concepts, e.g. when ‘bar’ and ‘man’ are linked in the compound *barman* in such a way that a new concept (‘a man who serves drinks in a bar’) emerges. While it is certainly true that compounds such as *fog+horn* express a single idea (in this particular case one comparable to a siren), the opposite is not necessarily true, because a single idea can also be expressed by a construction which is very clearly a phrase, e.g. ‘the smell of fresh rain in a forest in the fall’ or ‘the woman who lives next door’ – for which the English language has no equivalent compounds (Plag 2003: 7). Since one may, however, agree that all compounds as ‘complex lexemes’ refer to a unified semantic concept, whereas the majority of English phrases do not represent a single idea, the ‘unified semantic concept’ test can be used to determine potential candidates for compoundhood and will therefore be included in the final definition (cf. 2.6). For grammatical compounds such as *without*, the analogical requirement is a unified syntactic function, reflected in the assignment of a joint part of speech.

To determine in practice whether a construction in an English text represents a unified semantic concept, the interruptability test (cf. 2.1.2.4) can be carried out. If a sequence can only be interrupted with a change in meaning, it is considered to refer to a single idea. This test requires a very high command of English, and an inverse correlation between the number of constructions accepted as compounds and a speaker’s level of English can be expected, because the failure to imagine possible interrupting items may prompt less advanced speakers to classify borderline cases as compounds.

2.2 **Compounds versus Other Lexemes**

Besides the distinction from phrases on the syntactic level, compounds need to be set apart from other lexemes in the morphological dimension.

2.2.1 *Compounds versus Simplex Lexemes*

In the majority of cases, the distinction between compounds and simplex lexemes should not pose any practical problems, as a simplex like *tree* will rarely provoke uncertainty regarding potential categorisation as

a compound. However, there are two types of exception: on the one hand, so-called *fossilised compounds* (Dressler 2005: 40) or *obscured compounds* (Götz 1971) are no longer recognisable as compounds, so that e.g. *lord* and *lady* cannot be analysed into constituents in present-day English anymore. As a consequence, one may argue in favour of their classification as simplex lexemes from a synchronic perspective. Conversely, unmotivatable but transparent lexemes (cf. Sanchez 2008: 87) could theoretically be analysed into free pseudo-constituents, e.g. *forget* into *for+get*. This is for example the case of *mush+room*, a popular etymological interpretation of the French loanword *mousseron* (cf. *Oxford English Dictionary* [OED] s.v. *mushroom*). For the purposes of the present study (cf. also 4.1), the concept of the compound is therefore restricted to compounds which can be analysed into motivating – and thus semantically relevant – constituents in present-day English. While excluding pseudo-analyses such as *forget* and *mushroom*, the present approach includes all compounds whose parts appear morpho-semantically relevant, without consideration of their actual etymological origin.

2.2.2 *Compounds versus Derivatives*

In the majority of cases, the distinction between compounds and affixations is relatively unproblematic: compounds (e.g. *pen friend*) are formed from freely occurring constituents, and derivatives (e.g. *befriend* or *friendship*) contain at least one bound lexical affix. However, the definition of compounds cannot rely solely on the absence of lexical affixes, since some compounds (e.g. *ozone-friendly*) contain prefixes or suffixes within the compound's constituents. Yet this is not the only problem with regard to the dividing line between compounding and derivation:

1. The status of a particular morpheme as a prefix is not always easy to determine, because some prefixes (e.g. *after-*) are formally and semantically identical with free morphemes (in this case, the preposition *after*). For the sake of consistency, constructions containing the affixes listed in Table A.8 on the highest level of analysis are therefore considered prefixations rather than compounds in the present study (unless the meaning of the morpheme in question is different from the meaning of the affix, e.g. in the case of *postcard* or *adland*, which clearly refer to mail and advertising rather than to the meanings of the listed affixes).
2. Compounds containing one or more combining forms of Greek or Latin origin (Plag 2003: 74), such as the neoclassical compound

biology, are special in that their constituents cannot occur as free lexemes in their own right. Since this contradicts the requirements of the compound definition at the beginning of this chapter, such items are not considered compounds in the present study.

3. The same is true of lexemes of the type *cranberry*, which contain morphs that are unique in the language (cf. Aronoff 1976: 15). Although *berry* is a noun in its own right, *cran* never occurs on its own with the meaning it has within the complex word (i.e. not the homonymous Scotch form of *crane* recorded in the OED). Once again, this contradicts the vital requirement for compound status, “the declaration of independence” of a construction’s constituents (Bauer 2005: 97).
4. Compound-final *man* forms a large number of compounds with solid spelling (e.g. *fireman*, *policeman*) and has relatively general semantics, with its meaning corresponding to little more than the suffix *-er*. It can therefore be considered an *affixoid* (cf. Bußmann 2002 s.v. *Affixoid*, *Suffixoid*), but since the important difference to suffixation is that *man* also occurs as a free lexeme in its own right, such constructions are regarded as regular compounds here.
5. The absence of lexical affixes on the highest level of analysis is a particularly important criterion for the distinction between compounds and affixations. It plays a role in the categorisation of a number of borderline cases, particularly constructions ending in *-ed* or *-er* (e.g. *bite-sized* and *do-gooder*). Since *bite* and *sized* occur as lexemes in their own right in the English language and are also listed in the OED, in contrast to the verb *to bitesize* (from which *bite-sized* could have been derived), the complex adjective *bite-sized* is conferred compound status. The potential constituent *gooder*, by contrast, does not seem to exist as an English word and is not listed in the OED, which makes the derivation of *do-gooder* from the existent phrase *to do good* by means of suffixation the more plausible choice. However, even if all constituents of a potential compound exist as individual words (e.g. *black* and *marketeer*), the semantics of their combination need to be taken into account: since *black marketeer* does not refer to a marketeer with black skin but to someone selling objects on the black market, it is considered a suffixation of a phrase, i.e. [*black market*] + *-eer*, and not a compound. In the literature, the term *synthetic compound* is sometimes used in order to accept as compounds constructions which combine an initial noun with a final deverbal noun whose status as a word in its own right is doubtful: thus *goer* in *church+goer* and *swallower* in *sword+*

swallower are “possible, but not established English words”, which “function as building blocks in word-formation” through the simultaneous application of derivation and compounding (Booij 2007: 92).

Usually, derivatives are spelled solid, which should simplify the discrimination between compounds and derivatives, but there are some exceptions: we find hyphenation in some prefixations (e.g. *co-operate* or *re-cover*), particularly if there are orthographic or semantic reasons, such as the avoidance of sequences of identical letters or an unusual meaning (cf. 5.1.1), and also in some suffixations (e.g. *bell-less*, *shell-like*; cf. Ritter 2005a: 56) – but no spacing.

2.2.3 *Compounds versus Other Word Formations*

In most cases, the difference between compounds and other types of word formation beside derivatives is relatively clear, as demonstrated by these classical examples:

- Compounds differ from acronyms in that the latter (e.g. *BBC*) are usually capitalised throughout and consist of constituents which have been shortened so extremely (*British Broadcasting Corporation*) that they cannot be considered recurring free lexemes anymore.
- Blends such as *brunch* also contain constituents that have been shortened in such a way that they do not meet the requirement of representing recurring free lexemes (*breakfast lunch*).
- Clippings (e.g. *lab* from *laboratory*) as shortenings are unlikely to be confused with the more complex compounds. When clippings occur as parts of potential compounds, however (e.g. *language+lab*; cf. 2.5), one may disagree whether such constructions are better classified as compounds consisting of two free constituents (*language* and *lab*) or as clippings of a longer compound (*language+laboratory*). Whenever the clippings in such constructions recur as free words in English, the present study assigns compound status to them, because they fulfil the requirements of the compound definition (cf. 2.6).
- Back-formations (e.g. *to edit* from Latin *editor*; cf. OED) are not necessarily complex – but when they are, e.g. in the case of *to baby-sit* (from *baby-sitter*), the line is difficult to draw. Since they consist of free lexemes and have no affixation on the highest level of analysis, back-formations are not singled out by the compound definition used here (cf. 2.6). However, that is not necessary either, because back-formations are not distinct from compounds in their structure synchronically; only

historically (Huddleston and Pullum 2012: 286). A synchronic account of compounding therefore permits a certain extent of overlap between the two categories. As long as the result is compatible with the compound definition used, a compound may have undergone various word formation processes (cf. also Huddleston and Pullum 2002: 1660).

- Conversions are frequently simple lexemes, e.g. the verb *to bottle*, which goes back to the noun *bottle* (cf. e.g. Sanchez 2008: 93–94). When phrases or sentences consisting of more than two components are transformed into a construction with a single part of speech (e.g. in *an I-don't-care-what-you-do attitude*), this could be regarded as an instance of conversion. The present study will, however, follow the relatively common view that hyphens or concatenation in such constructions are an indication of a unified idea (cf. also Schmid 2011: 122) and categorise them as phrase compounds (cf. 2.5.2, 5.2 and Huddleston and Pullum 2002: 1660).

2.3 Compounds versus Multi-word Items

Since compounds consist of two or more free constituents, they need to be considered in relation to multi-word items, which always use open spelling. Jackson and Amvela (2002: 63–64) categorise compounds as one of the three main types of multi-word lexeme beside multi-word verbs and idioms.

Within the **multi-word verbs**, Quirk et al. (1985: 1150) distinguish phrasal verbs (*bring up; sit down*), prepositional verbs (*call for; look at*) and phrasal-prepositional verbs (*check up on; get away with*). While early English grammars often treat phrasal verbs as compounds and occasionally use hyphenation (e.g. *came-in* and *takes-away* in Solomon Lowe's 1737 *English Grammar Reformd* ...; cf. Sundby 1997: 227), phrasal verbs are generally considered syntactic phenomena and exclusively spelled open in present-day English. The present study follows this established convention and excludes verbal combinations of verb and adverb and/or preposition from the category of compounds.

The category of **idioms** can be considered as consisting of “grammatical units larger than a word which are idiosyncratic in some respect” (Croft and Cruse 2004: 230). Some of the various existing idiom definitions (cf. Croft and Cruse 2004: 230–236) are gradient towards compounding, e.g. Cruse's (1986: 37) requirement for idioms to “consist of more than one lexical constituent” and to represent “a single minimal

semantic constituent". The presumably most common definition is that idioms are constructions whose meaning cannot be predicted from the meaning of the several orthographic words composing them (Palmer 1981: 36; cf. also Lipka 2002: 90), e.g. *to kick the bucket*, *to bury the hatchet* or *to let the cat out of the bag* (Jackson and Amvela 2002: 65–66). Although idioms in this sense represent a semantic unit, they do not necessarily function grammatically like one: in contrast to the majority of compounds, idioms do not add inflection at their end. There is thus no past tense ²*kick the bucketed* (Palmer 1981: 80) – as against the past tense *freeze-dried* of the compound *freeze-dry* (whose first element is verbal and could carry inflection if the construction were a phrase). In view of the co-hyponymy of idioms and compounds, constructions which are sometimes referred to as idioms (e.g. a *red herring*) but take final inflection are considered compounds in the present study, while not excluding that they can be idioms at the same time.

Two additional types of multi-word item are discussed by Moon (1997: 45–47), namely fixed phrases (*of course*; *at least*; *in fact*; *how do you do*; *excuse me*; *you know*) and prefabs (*the thing is that*). These lexical chunks seem to have a very strong pragmatic function, particularly in spoken language, where their joint storage in the mental lexicon may save processing time.

Those **fixed phrases** which are formally whole sentences including a verb (e.g. *How do you do?*) and cannot be categorised as a single part of speech are not classified as compounds here. That the dividing line for the shorter fixed phrases is much more difficult to draw is reflected in their varying lexicographic treatment. For instance, the combinations *in fact*, *by far* and *at least* cannot be found in the electronic LDOCE, but the OED lists them as phrases. As a consequence, frequently recurring combinations of grammatical words need to be considered with particular care (cf. also Quirk et al. 1985: 672–673). Since concatenation of such combinations has occurred in the past, e.g. in the complex preposition *into* (cf. OED s.v. *into*), that is frequently regarded as a compound, and since change is still in progress (thus *of course* is lemmatised in some reference works such as LDOCE but only listed as a subentry in the OED), frequently recurring combinations of grammatical words need to be treated following an item-based approach.

Erman and Warren (2000: 31) define **prefabs** as combinations of "at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization" and distinguish the following subtypes:

- lexical prefabs (idioms, compounds, habitual collocations, phrasal and prepositional verbs)
- grammatical prefabs (e.g. quantifiers such as *a great deal of*; determiners such as *that sort of*; tense such as *be going to*; introductors such as *there is*)
- pragmatic prefabs (e.g. discourse markers such as *and then*; feedback signals such as *yeah quite*; hedges such as *I should think*)
- reducibles (e.g. *it's* or *I'm*) as a category that they consider more debatable.

While this very general definition of prefabs results in a very heterogeneous group of types, the distinction of prefabs from compounds is simplified by the fact that compounds are explicitly classified as a subcategory of prefab by Erman and Warren (2000). Since they define words orthographically, “*teacup* spelt as one word would not be considered a prefab, but *tea cup* spelt as two words would” (Erman and Warren 2000: 32). If one follows their definition, compounds should be considered prefabs, provided that they are spelled open and are sufficiently frequent. Based on that criterion, all open compounds listed in dictionaries can be considered prefabs – but not ad hoc compounds or relatively unestablished compounds.

Yet another addition to the set of multi-word items is constituted by the category of **collocation**, which overlaps to a large extent with fixed phrases and prefabs. There are two main definitions of *collocation*: the quantitative Firthian definition “[y]ou shall know a word by the company it keeps” (Firth 1957: 11) is based on lexical items’ high frequency of co-occurrence (cf. McEnery, Xiao and Tono 2006: 82) and purely statistical. Hausmann’s (1985: 118) qualitative approach, by contrast, posits that a collocation such as *to take a shower* is a typical, specific and characteristic relation between a relatively context-independent base (*shower*) and a collocator (*take*) that can only be understood in relation to a particular base (cf. Hausmann 2004: 311–312). Within both approaches to collocation, compounding may be considered a special type of collocation: thus Quirk et al. (1985: 1537) refer to low-frequency compounds as a “collocation [that] seems relatively unestablished” and Hausmann (2004: 317) states that a subset of compounds can be interpreted as collocations: thus the base *Dach* ‘roof’ of German *Schiebedach* ‘sunroof’ is relatively straightforward and likely to be translated by a direct equivalent into other languages, whereas the collocator *schiebe(n)* ‘push’ is unpredictable and e.g. translated into French by the meaning component ‘open’ in French *toit ouvrant* (literally, ‘opening window’) and the even more distant *sun* expressing the concept in English *sunroof*. As a consequence, the present approach considers compounds part

of the hyperonymous category of collocation, with frequency playing no clear delimitating role (as in the case of prefabs).

The longest type of multi-word item is represented by **proverbs** (Schmitt 2000: 99), e.g. *Out of sight, out of mind*. Since they represent full sentences ending with a punctuation mark (usually a full stop) and cannot be assigned a single part of speech, they are very clearly different from compounds.

To sum up, there is no general consensus regarding the relation between compounding and phraseology in the literature. In their comparison between the two, Granger and Paquot (2008: 32–33) – who also provide a detailed categorisation of phraseological units (Granger and Paquot 2008: 42–44) – find that phraseological approaches differ in their inclusion of compounds, with the traditional view tending to exclude either all or most compounds, whereas even hyphenated compounds may be categorised as multi-word units by alternative approaches. For a summary of the approach used in the present study, cf. 2.6.

2.4 Compounds versus Names

Names are a very special and easily recognisable category: they are always nouns, they are always capitalised and they refer to individual entities or people rather than having a generic meaning, the widely accepted view being that names “may have reference, but not sense” (Lyons 1977: 219). Since names may e.g. combine with other names to form longer names (e.g. two first names, such as *Mary Jane*, or a first name and a second name, e.g. *Elvis Presley*), the question arises whether such combinations should be considered compounds.

An important argument for the consideration of complex names as compounds is that their constituents may occur on their own and recur in the language. In combinations of first names, all three main compound spelling variants can be observed: there is open spelling (*Mary Jane*), hyphenation (*Mary-Lou*) and solid spelling (*Maryanne*). All three variants may coexist for a particular complex name (*Mary Lou*, *Mary-Lou* and *Marylou*), but in spite of the general liberties regarding the form of names, certain spellings seem to be avoided, e.g. solid *Malcolmchristopher*, which suggests that the principles applying to compounds (e.g. length, frequency etc.) may also be at work here. If two second names are combined (e.g. *Henderson-Smith*; cf. Morton Ball 1939: 30), both open and hyphenated spellings are possible in English (Soanes 2011).

Combinations of first and second names, such as *Malcolm Jones*, only seem to use open spelling – which makes them similar to phrases.

Yet an informal query among four native speakers of different varieties of English suggests that the relation between combined first and second names only superficially resembles that between the constituents of determinative compounds: someone who failed to understand a famous actor's first name in a conversation about films might ask: "Sorry, which Douglas did you mean? Kirk or Michael?", but an analogous question with compounds would be "Sorry, which room did you mean? Bedroom or bathroom?" rather than [?]"Sorry, which room did you mean? Bed or bath?" Furthermore, it is possible to ask for a surname ("Which Michael did you mean? Douglas or Jackson?") but not for a compound's head in a parallel question ([?]"Which bath did you mean? Room or tub?"). In view of this difference in structure and the fact that "either noun in a personal name may be used alone to indicate the person (object) referred to" in more or less intimate and symmetrical ways, the present study follows Morton Ball (1939: 30) in her classification of combinations of first and second name as appositions, which are typically noun phrases with reference identity (e.g. *Anna* and *my best friend* in the apposition *Anna, my best friend, was here last night*; cf. Quirk et al. 1985: 1300–1301). Names do thus not represent a unified category, and a distinction is made here between appositional combinations of first and second names and the combinations of two first names or two second names, which are considered compounds if the usual conditions for compound status apply (cf. also Morton Ball 1939: 30).

2.5 Compound Types

Much of the confusion regarding the delimitation of the category of compounds may be due to the fact that it potentially involves a large number of very different subtypes. The following sections present the various types which are commonly discussed in the literature, based on length, word formational structure, part of speech and spelling. Most accounts of compounding focus on noun+noun and noun+adjective constructions and otherwise frequently restrict themselves to "some major patterns" (Biber et al. 1999: 325). The following overview, by contrast, also attempts to provide a detailed discussion of minor compounding patterns, even though exhaustiveness may not have been achieved. Where the status of a category is controversial, the present approach follows Bauer's (1998: 65) tradition of the "lumper" by accepting as compounds all categories classified as such in the

literature, as long as they are compatible with the preliminary compound definition given earlier.

2.5.1 Length

Many English compounds – in any case the most typical ones – contain only two constituents (Quirk et al. 1985: 1567; Schmid 2011: 121), but each constituent of a determinative compound (cf. 2.5.2) may itself be a compound, e.g. *[[motor + cycle] + [outlet + store]]*, and the existence of longer compounds seems to be generally accepted in the literature. Coordinative compounds (e.g. *secretary-treasurer-editor* or *Metro-Goldwyn-Mayer*; cf. Huddleston and Pullum 2002: 1648 and 2.5.2) can also consist of more than two constituents, and phrase compounds such as *love-lies-bleeding* (cf. 2.5.2) even have more than two constituents by definition. According to Quirk et al. (1985: 1567), English compounds can involve any number of constituents, but while there is no clear cut-off point, readers seem less prepared to accept a construction as a compound with growing complexity (Donalies 2003: 78). As a consequence, some scholars may disagree with Plag's (2003: 133) example *university teaching award committee member* or Adams' (2001: 79) claim that *UK film industry task force appointment controversy* represents a compound.

2.5.2 Structure

Table 2.2 summarises the various compound categories recognised in the present study from the perspective of compound structure. It draws on Adams (2001: 82), Bauer (1983: 30–31, 212–213, 233), Marchand (1969: 11–127, 380–389) and Quirk et al. (1985: 1570–1578).

The status of phrase compounds is frequently disputed in the literature (e.g. by Meibauer 2003: 185; Adams 2001: 3; Plag 2003: 136), because they are supposedly not formed according to the usual rules of word formation and rather similar to conversion (Quirk et al. 1985: 1563, 1569). However, other accounts (e.g. Bauer 1983: 206–207) do recognise them as a type of compound. Among the small number of established phrase compounds, there are many family terms ending in *-in-law*, plant names (e.g. *forget-me-not*) and coordinated constructions (e.g. *bread-and-butter*; cf. Schmid 2011: 133–134). Since phrase compounds in general do not contradict the compound definition adopted here (cf. 2.6), and since some phrase compounds cannot be syntactic

Table 2.2 *Compound types based on word formation*

Name(s)	Characteristics	Example
Endocentric compound Determinative compound	The first element modifies the second (the head) The compound is a hyponym of the head	<i>armchair</i> : an armchair is a kind of chair
Exocentric compound Bahuvrihi compound	The compound is a hyponym of an unexpressed semantic head	<i>highbrow</i> : a highbrow is not a kind of brow but a kind of person
Appositional compound Copulative compound	The compound is a hyponym of both constituents	<i>maidservant</i> : a maidservant is both a type of maid and a type of servant
Dvandva compound	The compound is not a hyponym of either constituent The constituents denote separate entities which combine in an AND-relation to form the entity denoted by the compound	<i>Alsace-Lorraine</i> : Alsace- Lorraine consists of two separate entities, Alsace and Lorraine
Phrase compound	The compound contains at least one phrase or clause	<i>easy-to-read</i> (adj): consists of an adjective and a postmodifying <i>to</i> - infinitive (clause)
Genitive compound	The compound consists of a lexicalised noun phrase with a first constituent in the genitive, which acts as a linking morpheme like the German <i>Fugenelemente</i>	<i>bull's-eye</i> : <i>bull</i> + genitive <i>s</i> + <i>eye</i>

due to their ill-formedness (cf. the missing article in *a pain-in-(the-)stomach gesture*; Bauer 1983: 207), phrase compounds are also recognised as a subcategory of compounds in the present study.

The status of genitive compounds such as *bull's-eye* (cf. also 5.1.3.1 and 5.5.1.2) is by no means undisputed, either (cf. Sauer 1985: 309): they are frequently denied compound status (cf. Bauer 1983: 240–241) due to the fact that their constituents are linked by inflection, which makes them relatively phrase-like. However, interruption and reordering may result in semantic changes compared to phrases, e.g. if the compound *bull's-eye* ‘target’ is contrasted with the modified phrases *a bull's blue eye* and

the eye is a bull's. As a consequence, the approach adopted here follows Adams (2001: 80), who recognises a compound category “[n]oun-genitive *s* + noun”, whose members need not necessarily be as idiomatic as the example given earlier, e.g. *potter's wheel*.

By contrast, a number of categories considered compounds by other accounts of English compounding (cf. the sources of Table 2.4) were not generally included, as they might contradict the preliminary definition:

- **Neoclassical compounds** or **combining form compounds** such as *biology* consist of bound roots (*bio-*, *-logy*) by definition, which contradicts the requirement of freely occurring constituents for the whole category.
- **Rhyme-motivated compounds** consist of two rhyming elements (e.g. *hoity-toity*). According to Bauer (1983: 213), one of these may not exist independently, and judging from his examples, this may extend to both parts. While this contradicts the present approach's requirement of free occurrence, rhyme-motivated compounds are accepted when their constituents occur on their own, e.g. in *brain-drain*.
- **Ablaut-motivated compounds** consist of two elements differing only in their stressed vowel, e.g. *shilly-shally*. If at least one element does not occur freely (e.g. the first part of *wishy-washy*; cf. OED), they cannot be accepted as compounds here – but if both constituents do (e.g. *riff-raff*; cf. OED), such items are recognised as compounds.
- **Clipped compounds** contain one or more clippings, e.g. *optical art* or *situation comedy* (Bauer 1983: 233–237). Since the clipping must have taken place after the compounding process in the forgoing examples (cf. 2.2.3), they are not compounds according to the present approach. However, if the constituents of such constructions occur on their own, usually as informal clippings (e.g. in *language laboratory* or *parachute jump*), they are considered compounds here.

Last but not least, the existing compound categories from the literature could be complemented and/or refined by the following potential structural compound types:

- By analogy to Aronoff's (1976: 15) *cranberry morph*, the term *cranberry compound* could be coined to designate constructions such as *cranberry*, which contain constituents that are unique in the language but at the same time contrast with compounds containing two (or more) free constituents, e.g. *black+berry* or *blue+berry*. Since *cran* does not occur on its own in the language (except in linguistics texts pointing out this fact, as in the present sentence, which would skew any corpus search),

this category cannot be part of the compound definition used here but might complement other types of approach.

- The categories of copulative compounds and dvandva compounds could be further refined by considering some recurring but descriptively neglected patterns:
 - In dictionary titles, the order of the constituents expresses directionality: thus *English–Irish* is interpreted to refer to a dictionary providing translation equivalents of English words into Irish, whereas the opposite is the case of *Irish–English*. The relation expressed is thus ‘from... into...’.
 - In addition to the meaning component ‘against’, the order of the constituents in football matches expresses that the match takes place in the first of the two locations (e.g. Scotland in *the Scotland–France match*)
 - In the term “yes-no interrogative clauses” (Quirk et al. 1985: 724), *yes* and *no* are connected by an ‘or’ relation instead of the usual ‘and’ relation found in copulative and dvandva compounds.

2.5.3 Part of Speech

The classification of compounds according to the parts of speech involved in their formation is of particular interest with regard to spelling conventions. The overview in Table 2.4, which draws heavily on Bauer (1983: 201–216) and occasionally on Adams (1973, 2001), Aarts (2011: 34–35) and Tournier (1985: 113–119), seeks to provide examples for all three spelling variants when these are attested as examples in the reference works on word formation or in the dictionaries used in the present study (particularly MED and LDOCE; cf. 4.1). Table 2.4 shows the large variety of compound types in English. It lists compounds with a maximum of three constituents and thus more types than most other accounts of English compound formation but is still far from being exhaustive. Usually, a single example is given, except in order to draw attention to large differences between the examples. Additional compound types were added whenever the compounds encountered in the dictionaries could not be classified into any of the existing categories. The names of some categories (e.g. *gerund* + *n* or *n* + *deverbal noun*) were modified (in this particular case into *n-ing* + *n* and *n* + *n-ing*), and numbers were classified as numerals rather than adjectives. Since the part of speech of compound constituents is often difficult to determine, the empirical study described here uses a very limited set of

Table 2.3 *Labels used for compound classification in Table 2.4*

Label	Example	Explanation
active declarative clause	<i>love lies bleeding</i>	
adj	<i>new</i>	
adj- <i>ed</i>	<i>masked (ball)</i>	adjectives which are formally identical with a past participle – in compounds whose paraphrase tends away from a verbal meaning (not ‘the ball is masked’)
adv	<i>long(-playing)</i>	lexical adverbs (which may e.g. be formally identical with an adjective or are derived from adjectives by means of the suffix <i>-ly</i>)
det	<i>an</i>	correspond to the category <i>determiner</i> in Quirk et al. (1985)
imperative clause	<i>forget me not</i>	
interjection	<i>ha</i>	
n	<i>cable</i>	
n- <i>ing</i>	<i>(bicycle) repairing</i>	nouns which are formally identical with a present participle
noun phrase	<i>the cuff</i>	
n-proper	<i>Oxford</i>	
num	<i>two</i>	
particle	<i>out</i> <i>so</i>	grammatical words in several parts of speech: prepositions, conjunctions and grammatical adverbs
prepositional phrase	<i>about town</i>	technically the same as particle + n, but the label represents compound structure more accurately by indicating closer phrase-internal links between constituents
pron	<i>something</i>	
to-inf. clause	<i>to be</i>	
v	<i>meet</i>	
v- <i>ed</i>	<i>(drug-)related</i>	past participles of verbs – e.g. in compounds whose paraphrase has verbal meaning
v- <i>ing</i>	<i>(not-)withstanding</i>	present participles of verbs – e.g. in compounds whose paraphrase has verbal meaning
wh-clause	<i>what’s it</i>	

parts of speech, which subsumes all the grammatical parts of speech under a hyperonymous concept (cf. 5.6.1). In Table 2.4, however, an intermediate approach between the more traditional part-of-speech classification used for the whole compounds and that followed for the constituents in the empirical study is applied in order to do justice to the more detailed constituent-based compound type classifications found in the literature. Table 2.3 lists the parts of speech, phrases and clauses that are distinguished in Table 2.4. In addition, the genitive and the word *and* are used for recurring patterns.

Table 2.4 Compound types based on part of speech

PoS of compound	PoS of constituents	Examples		
		Open	Hyphenated	Solid
n	n + n	<i>cable television</i> <i>killer app</i> <i>MP3 player</i>	<i>meter-maid</i> <i>hunter-gatherer</i>	<i>manservant</i> <i>spoonbill</i>
n	n + n + n	<i>law enforcement agent</i>		
n	n + 's + n	<i>mama's boy</i>	<i>bull's-eye</i>	
n	n + n- <i>ing</i>	<i>night flying</i>	<i>bicycle-repairing</i>	
n	n- <i>ing</i> + n	<i>fishing rod</i>		
n	n + n-proper	<i>man Friday</i>		
n	n-proper + n	<i>Oxford accent</i>		
n	n-proper + n + n		<i>Wellington airport</i>	
n	n-proper + n-proper	<i>Mary Jane</i>	<i>Cadbury-Schweppes</i>	<i>Marylou</i>
n	n-proper + 's + n	<i>Adam's apple</i>		
n	n + <i>and</i> + n	<i>kith and kin</i>	<i>whisky-and-soda</i>	
n	n + adj	<i>heir apparent</i> <i>machine washable</i>	<i>knight-errant</i>	
n	n + adv	<i>centre forward</i>		
n	n + particle + n	<i>morning-after pill</i>		
n	n + particle		<i>looker-on</i> <i>passer-by</i>	<i>checkout</i>
n	n + prepositional phrase	<i>man of God</i> <i>man about town</i>	<i>mother-of-pearl</i>	
n	n + <i>to</i> -inf. clause		<i>mother-to-be</i>	
n	n + num	<i>number one</i>		
n	v + n	<i>install program</i>	<i>goggle-box</i>	<i>pickpocket</i>
n	v + v		<i>make-believe</i> <i>look-see</i> <i>has-been</i> <i>might-have-been</i>	<i>hearsay</i>
n	v + v + v			
n	v + <i>and</i> + v	<i>meet and greet</i>		
n	v + adv		<i>get-together</i>	<i>lookalike</i>
n	v + particle	<i>sod all</i>	<i>drop-out</i>	<i>cookout</i>
n	v + particle + n	<i>lighting up time</i> <i>jumping-off point</i>		
n	v + interjection		<i>heave-ho</i>	
n	adj + n	<i>new town</i>	<i>fast-food</i>	<i>software</i> <i>outerwear</i>
n	adj- <i>ed</i> + n	<i>masked ball</i>		
n	adj + adj		<i>creepy-crawly</i>	
n	adj- <i>ed</i> + pron	<i>loved one</i>		
n	adj + n + n	<i>hot cross bun</i> <i>hot-water bottle</i>		

Table 2.4 (*cont.*)

PoS of compound	PoS of constituents	Examples		
		Open	Hyphenated	Solid
n	adj + adj + n	<i>obsessive</i> <i>compulsive</i> <i>disorder</i>		
n	adj- <i>ed</i> + particle		<i>grown-up</i>	
n	adv + v		<i>also-ran</i>	
n	adv + v- <i>ing</i> + n		<i>long-playing record</i>	
n	adv + v- <i>ed</i> + n		<i>ill-gotten gains</i>	
n	adv + adj- <i>ed</i> + n	<i>less developed</i> <i>country</i>		
n	adv + particle		<i>close-up</i>	
n	pron + n	<i>It girl</i>	<i>she-goat</i>	
n	particle + n	<i>in box</i>	<i>in-crowd</i> <i>twice-winner</i>	
n	particle + v		<i>to-do</i>	
n	particle + particle		<i>once-over</i>	
n	prepositional phrase + n- <i>ing</i>	<i>in-line skating</i>		
n	num + n			<i>hundredweight</i>
n	num + pron			<i>twentysomething</i>
n	num + particle		<i>eleven-plus</i>	
n	num + num		<i>one-two</i>	
n	num + particle + num		<i>four-by-four</i>	
n	det + particle + n	<i>no through road</i>		
n	interjection + interjection		<i>hoo-ha</i>	
n	active declarative clause	<i>keep fit</i>	<i>love-lies-bleeding</i> <i>I-spy</i>	
n	imperative clause		<i>forget-me-not</i>	
n	<i>wh</i> -clause			<i>whatsit</i>
v	n + n + v	<i>crystal ball-gaze</i>		
v	n + v	<i>hand wash</i>	<i>lip-read</i>	<i>brainwash</i>
v	v + n			<i>leapfrog</i>
v	v + v	<i>dare say</i> <i>make do</i>	<i>trickle-irrigate</i> <i>stir-fry</i>	<i>typewrite</i>
v	adj + n		<i>bad-mouth</i>	<i>deadhead</i>
v	adj + v	<i>warm iron</i>	<i>free-associate</i>	<i>whitewash</i>
v	adv + v		<i>left-click</i>	
v	num + v		<i>second-guess</i>	
adj	n + n		<i>king-size</i>	<i>borderline</i>
adj	n + <i>and</i> + n		<i>meat-and-potatoes</i>	
adj	n + v- <i>ing</i>		<i>ocean-going</i>	
adj	n + v- <i>ed</i>		<i>drug-related</i>	

Table 2.4 (*cont.*)

PoS of compound	PoS of constituents	Examples		
		Open	Hyphenated	Solid
adj	n + adj	<i>medium dry</i> <i>HIV positive</i>	<i>lime-green</i>	<i>childproof</i>
adj	n + particle		<i>bottom-up</i>	
adj	n + prepositional phrase		<i>matter-of-fact</i>	
adj	v + n		<i>roll-neck (sweater)</i>	<i>breakneck</i>
adj	v + v		<i>stop-go (economics)</i>	
adj	v + <i>and</i> + v	<i>nip and tuck</i>	<i>kiss-and-tell</i>	
adj	v + adj		<i>feel-good</i>	
adj	v + adv		<i>go-ahead</i>	
adj	v + particle		<i>see-through (blouse)</i>	
adj	adj + n		<i>red-brick (university)</i>	<i>wholemeal</i>
adj	adj + adj		<i>deaf-mute</i>	
adj	adj- <i>ed</i> + particle	<i>messed up</i>	<i>hoped-for</i>	
adj	adj + particle	<i>high up</i>		
adj	adj + prepositional phrase	<i>hard of hearing</i>	<i>honest-to-goodness</i>	
adj	adv+ adj- <i>ed</i>			<i>newborn</i>
adj	adv + v		<i>long-stay</i>	
adj	adv + v- <i>ed</i> + particle		<i>long-drawn-out</i>	
adj	adv + adv			<i>faraway</i>
adj	adv + particle		<i>far-off</i>	<i>nearby</i>
adj	adv + prepositional phrase		<i>just-in-time</i>	
adj	pron + adv		<i>me-too</i>	
adj	prepositional phrase	<i>in depth (study)</i>	<i>before-tax (profits)</i> <i>off-the-cuff</i> <i>off-centre</i>	<i>indoor</i>
adj	particle + n			
adj	particle + v- <i>ing</i>			<i>ongoing</i>
adj	particle + adj		<i>all-important</i>	
adj	particle + particle		<i>on-off</i>	
adj	particle + <i>and</i> + particle		<i>out-and-out</i>	
adj	num + n		<i>five-star</i> <i>16th-century</i>	
adj	num + adj- <i>ed</i>		<i>one-sided</i>	
adj	num + adv		<i>first-ever</i>	
adj	num + particle + num		<i>nine-to-five</i> <i>one-on-one</i>	
adj	num + particle		<i>one-off</i>	

Table 2.4 (*cont.*)

PoS of compound	PoS of constituents	Examples		
		Open	Hyphenated	Solid
adj	num + num		<i>fifty-fifty</i>	
adj	det + n		<i>no-nonsense</i>	
adj	det + v		<i>no-go</i>	
adj	active declarative clause		<i>one-size-fits-all</i>	
adv	n + particle	<i>inside out</i> <i>head first</i>		
adv	v + v			<i>maybe</i>
adv	adv + adv		<i>double-quick</i>	
adv	adv + particle	<i>high up</i>		<i>nearby</i>
adv	adv + det+ adj			<i>nevertheless</i> <i>nonetheless</i> <i>indeed</i>
adv	particle + n			
adv	particle + prepositional phrase		<i>up-to-the-minute</i>	
adv	prepositional phrase	<i>of course</i>	<i>off-the-record</i>	<i>indoors</i>
adv	num + num		<i>fifty-fifty</i>	
adv	det + n	<i>no place</i>		<i>meanwhile</i>
adv	det + particle			<i>nowhere</i>
prep	adv + particle			<i>nearby</i>
prep	particle + v- <i>ing</i>			<i>notwithstanding</i>
prep	particle + particle	<i>because of</i>		<i>onto</i>
pron	pron + n			<i>somebody</i>
pron	pron + pron	<i>one another</i>		<i>anyone</i>
pron	det + n			<i>nobody</i>
pron	det + pron	<i>no one</i>	<i>no-one</i>	
conj	particle + adv			<i>whenever</i>
conj	particle + particle	<i>so that</i>		
num	num + adj		<i>80-odd</i>	
num	num + pron			<i>twentysomething</i>
num	num + num	<i>two hundred</i>	<i>twenty-two</i>	
det	det + det			<i>another</i>
interjection	n + n			<i>fiddlesticks</i>
interjection	interjection + interjection	<i>tsk tsk</i>	<i>uh-oh</i>	
interjection	declarative clause	<i>thank you</i>		
interjection	imperative clause			<i>farewell</i>

Note that in Table 2.4, the simultaneous presence of open and solid spelling (e.g. in the triconstituent compound *Wellington airport*) is placed in the middle of the scale for want of a better location, whereas combinations of hyphenation with one of the other two types are situated in more iconic positions.

The listing of part-of-speech-based compound types in Table 2.4 follows Marchand's (1960a: 20) view that "[c]ompounding occurs in all word classes" by including compound verbs (which are e.g. not considered by Quirk et al. 1985), compounds formed from grammatical morphemes (which are denied compound status e.g. by Schmid 2011: 128) as well as compound adverbs, determiners, numerals and interjections. However, the comparison between the compounds found in the literature with the parts of speech in Quirk et al. (1985: 67–77) suggests that some parts of speech are never compounded, namely modal verbs, primary verbs and the members of the category 'unclassified' (i.e. the negative particle *not* and the infinitive marker *to*). Nonetheless, patterns which are missing from Table 2.4 are not automatically impossible (for a discussion of potential word formations, cf. e.g. Burgschmidt 1977; Bauer 2001): as a considerable number of patterns in Table 2.4 are based on the empirical study's compounds with identical spelling in five to six dictionaries, future research is likely to yield even more compound types to add to the list.

2.5.4 Spelling

When separate lexemes or grammatical words are combined into compounds, a distinction has to be made in writing which is not required in oral speech (cf. Quirk et al. 1985: 1614): while varying vowel and consonant length, different degrees of loudness or pauses of different length may blur word or constituent boundaries in spoken language (e.g. between *a nice drink* and *an ice(d) drink*; Quirk et al. 1985: 1614), the "visual indicators of word limits" permit no gradience, and writers of English are usually forced to make an absolute decision between "total separation, hyphenation, and total juxtaposition" (Quirk et al. 1985: 1614), the three main types of English compound spelling.

The constituents of so-called **open compounds** (Merriam-Webster 2001: 99) are separated by one or more spaces, depending on the number of constituents (e.g. *cable television* and *central nervous system*). Alternative terms are *spaced compound* (Juhasz, Inhoff and Rayner 2005) or *separated compound* (Huddleston and Pullum 2002: 1759–1760), the first of which seems to be more common. Since – as we have seen – the “status of spaced two-word segments is uncertain” (Sundby 1997: 225), some accounts of English compound spelling (e.g. Morton Ball 1951: 3) and some dictionaries from the nineteenth and early twentieth centuries denied compound status to open sequences (cf. Morton Ball 1939), but most present-day treatments of compounding accept and include open compounds.

In **hyphenated compounds** (Merriam-Webster 2001: 99), the constituents are separated by one or more hyphens, depending on the number of constituents (e.g. *hunter-gatherer* and *forget-me-not*). The alternative terms *hyphenated compound*, *hypheme* (Morton Ball 1951: 3), *half-compound* and *occasional compound* (Sundby 1997: 225) seem to be uncommon. While some psycholinguistics studies (e.g. Juhasz et al. 2005) restrict compound spelling to the distinction between open and solid spelling and do not consider hyphenated compounds, this is relatively unusual.

Solid compounds (Merriam-Webster 2001: 99), such as *software* and *twentysomething*, constitute orthographic words and thus uninterrupted sequences of letters (cf. Plag 2003: 4). They are the only type of compound that is not excluded by any compound definition. Strumpf and Douglas (1988: 52) use the alternative term *closed compound*, which is common compared to less established alternatives such as *juxtaposed compound* (Huddleston and Pullum 2002: 1759–1760), *solideme* (Morton Ball 1951: 3) or *absolute compound* (Sundby 1997: 225).

Of the three spelling variants, solid spelling is the only one that can be considered immaterial. The hyphen obviously has a form, and although the gap in open compounds is not filled, it occupies a space which could have been used otherwise. The three types of compound spelling in English make use of the application or non-application of two principles:

- a) concatenation vs. non-catenation
- b) use of a hyphen vs. non-use of a hyphen.

The combination of these two options with their two possible values (i.e. the presence or absence of each feature) yields the three most common types of compound spelling in English, namely

1.	Open	(= non-concatenation + non-use of a hyphen),	e.g. <i>boy friend</i>
2.	Hyphenated	(= concatenation + use of a hyphen),	e.g. <i>boy-friend</i>
3.	Solid	(= concatenation + non-use of a hyphen),	e.g. <i>boyfriend</i> .

A fourth variant can be derived from the logical combination of these parameters, namely non-concatenation + use of a hyphen, e.g. *boy-friend*. While this variant does not seem to occur on its own in the English language, there are contexts in which it may be used, namely in combinations of identically structured hyphenated compounds, the first of which is incompletely realised on the formal level because of the ellipsis of the shared second constituent, e.g. in *car-owners and ship-owners* (Morton Ball 1939: 96) or in *boy- and girl-friends*. Longer sequences are conceivable, with the proviso that the shared second constituent is always retained in the last of the compounds (cf. Morton Ball 1939: 96), e.g. in *These are paragraph-, sentence-, or clause-boundaries* (Nunberg 1990: 69). However, this phenomenon, which is termed “floating hyphens” by Butcher (1992: 154) and “elliptical compounds” by Morton Ball (1939: 96), is frequently regarded as undesirable (cf. Quirk et al. 1985: 1347), and Butcher (1992: 154) suggests its avoidance by rewording.³

While the present study focuses on the three most common ways of spelling English compounds, alternatives using punctuation marks other than the hyphen exist or existed in the past:

- Between the fourteenth and eighteenth centuries, the **equals sign** < = > was sometimes used instead of a hyphen, at least in end-of-line hyphenation (McDermott 1990: 13). Especially in the United States, it is still used by proofreaders to “signify the instruction to insert a hyphen” and therefore now corresponds to an alternative variant with a different pragmatic focus (Clark 1990: 196).

³ In contrast to English, where the fourth spelling variant is only permitted in combinations of hyphenated compounds, and usually only when the last constituent is shared (Morton Ball 1939: 96), German more commonly applies it to originally solid compounds (which are the norm in German) and even permits the elision of the first part of the compound, e.g. in *Schuljungen und -mädchen* (Duden 2006: 1212), which would correspond to *schoolboys and -girls*. The only two English examples found in the literature for this phenomenon are *Australian-born and -educated* (Huddleston and Pullum 2002: 1761) and *type-setting or -casting machines* (Carney 1994: 49), which result in the otherwise unexpected possibility of encountering a hyphen line-initially. The omission of both a compound-final and an initial constituent in an enumeration, as in German *Textilgroß- und -einzelhandel* = *Textilgroßhandel und Textileinzelhandel* (Duden 2006: 1212), does not seem possible in English.

- A conventional (but not extremely frequent) alternative to the hyphen is the *oblique stroke* or *slash* < / >. Since slashes are not normally flanked by spaces (although usage may vary), Huddleston and Pullum (2002: 1731) define them as part of word-level punctuation, i.e. “the marking of word boundaries and the use of punctuation marks . . . within a word”. Slashes are used in very specific types of compound, e.g. some noun+noun compounds describing a double title or function, such as *bar/restaurant* (Merriam-Webster 2001: 100–101). Quirk et al. (1985: 1570) extend their use to coordinate compounds without restricting part of speech and give adjectival examples, e.g. *aural/oral (approach)*. Interestingly, the slash has an additive function rather than its usual alternative function in these compounds. In addition, the oblique stroke may be preferred over the hyphen “where one or more elements consist of more than one word, e.g. Bedford/Milton Keynes boundary” (Butcher 1992: 152).
- Single or double quotation marks may be used in the spelling of long English compounds (Morton Ball 1951: 6), e.g. those including a film title such as “*Gone with the Wind*” *remake* (Google search, June 2017). The pairwise occurrence of quotation marks (Meyer 1987: 4–6) opens a unified slot, whose closing is indicated by the second part of the pair. While quotation marks are clear delimiters of long components, they combine with spaces or hyphens and thus cannot be considered a completely new spelling variant.
- In programming languages, underscores are often used in order to connect the parts of a lexical entity, e.g. *window_id_format* (cf. Venezky 1999: 41).

From a systemic and logical point of view, any punctuation mark could be used to link the constituents of compounds – except those which usually indicate syntactic boundaries and need to be followed by a space. However, even though commas, full stops, colons, semi-colons, question marks and exclamation marks break up the unity of the compound, which would seem to prevent their use, there are some exceptions even here: thus the original compound list from the *Longman Dictionary of Contemporary English* (cf. 4.1) contained the two comma-separated items *all-singing, all-dancing* and *two up, two down*, and in the American spelling of *Mr. Right*, a compound-internal full stop occurs at the end of the abbreviated first constituent. Furthermore, in metalinguistic language use, compounds with undetermined spelling may be concatenated with a question mark between the constituents (*Rot?wein* ‘red?wine’; Jacobs 2007: 54).

Yet another possible means of marking constituent boundaries is the modification of the standard font: thus it is customary to italicise foreign phrases within compounds (cf. Fowler's 1921: 10 example "an *ex officio* member"), and superscript is occasionally combined with solid spelling when chemical elements and figures are conjoined (e.g. Sr^{90} ; *GPO Style Manual* 2008: 82).

2.6 Summary

The discussion and analysis of the spelling of English compounds are complicated by the fact that compounds represent a very heterogeneous category which seems to defy a general and generally accepted definition. As a consequence, research on English compounds is typically based on the respective scholar's own definition. Based on the previous sections' discussion of how compounds can be distinguished from syntactic constructions, other lexemes, multi-word items and names, the preliminary definition can now be refined: the present study defines English compounds as complex lexemes which

- refer to a unified semantic concept
- consist of at least two constituents that occur as free, synchronically recognisable and semantically relevant lexemes each
- contain no affixation on the highest structural level
- can be assigned a joint part of speech
- cannot be interrupted by the insertion of lexical material
- only once permit the application of each type of inflection to their base form

and (which is only important for specific subsets)

- do not follow the pattern 'name + *and/or* + name' or its variants with more constituents
- are not verbs consisting of a verb followed by an adverb and/or a preposition
- do not combine a personal name and another lexical entity with reference identity.

It is interesting to observe the large proportion of syntactic criteria used in the delimitation of compounds, which means that syntactic criteria are applied to delimit the boundaries of a particular type of lexeme in contrast to syntactic constructions – i.e. to distinguish compounds from phrases. Conversely, and as we have seen, lexical criteria (such as orthography and stress) may be used to distinguish phrases from compounds.

The detailed definition given earlier is unusual in its combination of very general principles to follow and very specific patterns to avoid. In spite of its relative precision, the definition cannot prevent a certain degree of overlap with other categories – as in other accounts of compounding, which usually conclude that there is a continuum from obvious compounds to obvious syntactic groups (cf. e.g. Schmid 2011: 133; Bauer 1998: 83; Mondorf 2009: 381; Quirk et al. 1985: 1570): while a small number of the adjacent categories discussed earlier (e.g. acronyms and proverbs) can be clearly set off from compounds, compounds cannot be distinguished from collocations, since the category of compounds is hyponymous to that of collocation and thus completely integrated into it. More commonly, however, there is partial overlap, e.g. with conversion: while most compounds are not conversions (e.g. the noun *bed+room*), and while many conversions are not compounds (e.g. the verb *to bottle*), the part-of-speech assignment in phrase compounds (e.g. the adjective *do-it-yourself*) can be likened to conversion. Furthermore, the presence of complex bases in shortenings (such as clippings or back-formations) may result in the emergence of borderline cases with simultaneous category membership (e.g. *language+lab* and *to baby+sit*), so that only simple shortenings can be set off from compounds. While the present study's wide compound definition results in borderline cases with gradience towards collocation or valency construction (e.g. the interjection *thank you* or the preposition *because of*), this does not affect the central empirical results, which are based on a sample of nouns, verbs, adjectives and adverbs and should therefore also be compatible with narrower compound definitions that e.g. do not accept grammatical compounds. Table 2.5 summarises the defining criteria for compounds. Those introduced by *because* apply to all compounds, whereas criteria introduced by *if* only apply to some compounds (e.g. those with fore-stress) when contrasted with a particular adjacent category.

A fundamental question with regard to the definition of the compound concept is whether there is an essence of compoundhood, from which all other criteria can be derived. A criterion which is so basic that it cannot be avoided is the exclusive composition from at least two free and recurring⁴

⁴ The decision when a potential compound's constituents can be defined as having existence in their own right is impossible to settle in a general way. In the present empirical study, existence is usually operationalised as listedness in the *Oxford English Dictionary* (OED 2009) as the largest dictionary of the English language. In addition, two adjectival, verb-derived forms which are structured by analogy to a multitude of other listed adjectival participle forms in *-ing* and *-ed* (such as *boggling* or *tied*) but unlisted in the OED for no obvious reason were also accepted as compound constituents: *finding* in *fact-finding* (adj) and *maintained* in *grant-maintained* (adj).

Table 2.5 *The delimitation of compounds*

Compounds differ from phrases	because they can be assigned a joint part of speech because they cannot be interrupted without resulting in a new compound with a different constituent structure (except superficially in conjoinths) because their base form permits the addition of each type of inflection only once (note the exception constituted by phrases consisting of name + <i>and/or</i> + name and its variants) if they are hyphenated or spelled solid if they have fore-stress if they are syntactically ill-formed
Compounds differ from simplex lexemes	because they contain more than one synchronically recognisable and semantically relevant constituent
Compounds differ from derivatives	because they only contain freely occurring constituents on the highest structural level of analysis
Compounds differ from acronyms	because their constituents are not shortened to individual letters
Compounds differ from blends	because their constituents are not shortened to non-recurring forms
Compounds differ from simple clippings	because they are not shortenings
Compounds differ from simple back-formations	because they are not shortenings
Compounds differ from conversions	if they are not phrase compounds
Compounds differ from phrasal verbs	because verbal constructions with open spelling combining a verb and an adverb are excluded by definition
Compounds differ from prepositional verbs	because verbal constructions with open spelling combining a verb and a preposition are excluded by definition
Compounds differ from idioms	because they either take inflection word-finally or on their initial head
Compounds differ from fixed phrases	if they are combinations of lexical words or shorter than a clause
Compounds differ from prefabs	if they are infrequent and either spelled solid or hyphenated
Compounds differ from proverbs	because they do not take the form of a sentence ending with a punctuation mark because they can be assigned a joint part of speech
Compounds differ from names	if they have generic meaning

constituents on the highest level of analysis. Another central starting point is the presence of a unified semantic idea (cf. also Lipka's 1977: 155, 161 and Schmid's 2008 discussion of *hypostatisation*), since the consideration of a construction's status as a compound would be futile if it lacked a unified semantic concept. Furthermore, compounds are inherently characterised by their function as single minimal syntactic units. While these basic requirements permit the direct derivation of some compound criteria (e.g. the assignment of a joint part of speech and the single addition of each type of inflection based on the unified syntactic function, or the interpretation of orthographic unity as a formal reflection of semantic unity), other compound criteria cannot be derived from the more basic criteria (e.g. fore-stress or syntactic ill-formedness). We can therefore distinguish between *defining criteria*, i.e. "conditions that have to be fulfilled" in order to achieve category membership, and *classifying criteria*, which are "applicable to varying degrees to different kinds of" category members (Handl 2008: 51). While all of the present account's defining criteria are listed in the compound definition given earlier, the other criteria discussed in this chapter (e.g. fore-stress, right-headedness or listedness) are optional and therefore only mentioned in the complementary category delimitation in Table 2.5 (if at all).

The difficulties in distinguishing compounds from other constructions can be attributed to the fact that there is often no reversible relation: thus criteria such as stress or spelling can frequently be used to delimit the contrasting category, e.g. that of phrases (with no fore-stress, solid spelling or hyphenation in syntactic constructions), but the criteria in question do not apply to all compounds, as some items considered compounds by the present approach and various other researchers (e.g. *apple pie*) have back stress and open spelling, like phrases.⁵ Nonetheless, many of these less distinctive criteria (including the unified semantic concept) can still contribute to the categorisation of individual items by means of clustering: the more compound criteria beyond the defining ones apply to a construction, the more indisputable the construction's status as a compound.

⁵ Similarly, while the word formation process of compounding always produces compounds, compounds may also be produced by other processes: according to Huddleston and Pullum (2002: 1646), "dephrasal compounds" such as *has-been* (n), *hard-core* (adj) and *cold-shoulder* (v) arise "not by the normal morphological process of compounding but rather through the fusion of words within a syntactic structure into a single lexical base".

The Normative Background

A comprehensive account of the spelling of English compounds also needs to examine the social factors that play a role in determining its shape. The following sections go beyond previous research by considering the issues of norm, prescriptivism and standardisation in relation to English compound spelling within their social background. Since English spelling norms are not fixed by any institution endowed with the legal power to prescribe linguistic usage, they emerge from within the community of language users.

Norms can be defined as “socially shared concepts of appropriate and expected behaviour” (Kauhanen 2006: 34) or as “negotiated behaviour regularities of social groups” (Lenz and Plewnia 2010: 11). Linguistic norms frequently correspond to the language of some group regarded as positive. While the prestige of a particular variant or variety is usually the most important factor for its selection (Milroy and Milroy 1985: 15), e.g. with regard to pronunciation of rhotic /r/, variation in English compound spelling does not seem to be loaded with any social meaning and does therefore not appear to act as a social marker.

Norms spread through language by the process of *standardisation*, which intends to achieve “the minimum of misunderstanding and the maximum of efficiency” (Milroy and Milroy 1985: 23). Milroy and Milroy (1985: 22) claim that English spelling has achieved full standardisation, which “*involves the suppression of optional variability*” (Milroy and Milroy 1985: 8), but as far as the spelling of English compounds is concerned, one may disagree.

An important distinction in this context is that between permissible variation and illicit mistakes. Defining a mistake is very difficult, since it needs to be distinguished from unusual spellings. Most linguists will presumably do this on the basis of frequency of usage, possibly also based on intentionality (cf. also Tavosanis 2007) and on whether errors are made consistently by a language user or constitute a momentary lapse in

performance (Carney 1994: 81). Furthermore, spellings are often assumed to be correct “if acknowledged as such by at least one authoritative source, such as a dictionary, even if other sources classify it as a misspelling” (Tavosanis 2007: 100). While the spelling of the possessive determiner *its* with an apostrophe (**it’s*) will presumably be criticised by all competent spellers, such a high level of agreement cannot usually be expected regarding the spelling of English compounds. As a consequence, a considerable number of English compounds can be classified as linguistic cases of doubt (*sprachliche Zweifelsfälle*; Klein 2003: 7), i.e. linguistic units in whose production competent speakers may feel in doubt as to which of two or more concurring and formally highly similar forms is correct in the standard variety. The difference between a linguistic “case of doubt” and mere instances of doubting about language is that the former always involves a plausible alternative (Dürscheid 2011: 159). This is always the case in compound spelling, since the three variants coexist as theoretically possible alternatives that are constructed following a systematic pattern. Linguistic cases of doubt differ from mistakes, since mistakes are judged as incorrect in hindsight by competent speakers, whereas linguistic cases of doubt remain unresolved and represent a more or less general problem in a community of speakers (Klein 2003: 7–8) – as is attested for English compound spelling by the manifold quotes in the Introduction.

3.1 Why ‘Correct’ Spelling Matters

Language users (not only of English) seem to feel very strongly that “there is a ‘right’ and a ‘wrong’ way to write words”, although that is “by no means a logical necessity” (Sebba 2007: 10). Even in cultures permitting deviations from an obligatory official standard in private writing, which would theoretically make the coding procedure much easier, this licence is rarely exploited (Sebba 2007: 44). This raises the question what makes users of a language follow particular spelling conventions. That spelling obviously matters, and that language users feel the need for the codification of certain practices, can be seen from the success of style guides such as Hart’s (1957) *Rules for Compositors and Readers*, which was originally only intended for staff at the Clarendon Press but met with more and more requests for copies from outside (Hart 1957: 5). While some language users settle doubtful spellings by using style guides or dictionaries, others carry out Google searches or trust their intuition – but one is unlikely to find language users who never reflect about spelling at all. In the majority of cases, users will try to comply with norms, even if these norms are only

implicit. The following sections discuss interrelated and partly overlapping reasons why 'correct' spelling matters. Most of these reasons are social rather than linguistic, because spelling (as part of language) is part of human social behaviour and consequently a type of social convention – one on which particular social groups may place more or less emphasis (Kress 2000: 26).

The main objective of using language is **successful communication**, and compliance with orthographic norms is beneficial to that aim: the amount of text that we read far outnumbers that which we write, and if all spellers follow orthographic norms, reading becomes much easier. While Grice's (1975: 45–46) cooperative principle, which postulates that speakers/writers contribute to conversation in such a way as is required by the hearer/reader and the situation, does not explicitly expect language users to use the most accepted linguistic form, Grice's maxims of manner suggest that speakers/writers tend to be brief and orderly and to avoid ambiguity and obscurity of expression (Grice 1975: 45). As standardisation contributes towards the clearness and ease of perception of a text, this may possibly explain language users' predisposition to observe orthographic norms. Even if the majority of divergences from standard spelling are presumably relatively unambiguous, inconsistencies such as different spellings of the same compounds within a short distance in one text (Butcher 1992: 64–65) may distract readers' attention away from the message – thus not impeding successful communication, but making it more difficult and thereby less economic.

Good command of spelling is also linked to issues of **power** and **control**, as the mastery of conventions may lead to pride (Ghameshi 2010: 9) and self-confidence. The downside of this aspect is that it may also give rise to a **sense of superiority** (Beal 2010: 63) and eventually to a situation in which the language of others is not only criticised to further what is regarded as linguistic correctness, but also as "a way of asserting superiority, judgment, and power over them" (Ghameshi 2010: 90).

Spelling "matters because **tradition** matters" (Ghameshi 2010: 89), and therefore failure to comply with the established norms – which convey a feeling of **stability** – is frequently likened not only to linguistic decline but also to moral decline (Allan and Burridge 2006: 122). Furthermore, the insistence on "there being a 'right' and 'wrong' way to use language serves to justify the way we have been taught" (Ghameshi 2010: 90), and the adherence to prescriptive orthographic principles could also be regarded as a way of **paying tribute to one's former teachers**. Since the mastery of correct spelling is frequently regarded as an **outward sign of education**,

linguistic correctness may also be considered an **immaterial status symbol**. After all, the rise of prescriptive grammars can be related to the wish of socially aspiring and linguistically insecure members of the middle class seeking advice on how to distinguish themselves from the speech of their social inferiors (Ghomeshi 2010: 73). In the twenty-first century, mastery of English spelling is not regarded as a special skill but rather taken for granted. As a consequence, failure to follow the commonly accepted norms may result in embarrassment (Antos 2003: 37) and a loss of **face**, because the interlocutors might attribute it to ignorance.

According to Sebba (2007: 160), “[t]he tendency of orthography to become a **marker of identity** is beyond question” – which might explain why orthographic issues frequently trigger extreme reactions. For instance, the linguist Kate Burridge received hate mail after suggesting that the English language could do without the hyphen in many contexts (Burridge 2010: 5). Ghomeshi (2010: 92) believes that “people are more likely to have a pet peeve about language than they are about other burdens of modern capitalism”, such as “illogical grocery store layouts”, because “there is a genuine interest in language and the ways it is used”. One reason for that interest may be that everyone uses language and therefore feels qualified to discuss it (Allan and Burridge 2006: 122). In spelling, as in other areas of language, linguistic means are used to signal identity whenever this is permitted through the existence of alternative choices. In orthographic identity construction, different forces are pulling in opposite directions: the wish to set oneself apart from others as a unique being – present, for instance, in the choice of unusual spellings like <Jessiqua> for children’s names (Ghomeshi 2010: 77) – as against the wish to fit into a group following certain traditional orthographic norms. With the rare exception of works of literature (cf. 5.10.2), the spelling of English compounds rarely seems to be employed in the first of these functions.

In contrast to identity (as the impression which language users consciously want to convey about themselves by employing particular linguistic features), **character** is what other language users try to infer from the linguistic features that a speaker/writer uses (Milroy and Milroy 1985: viii). The conclusions drawn from the degree of compliance with orthographic norms about the character of an unknown person (Figueredo and Varnhagen 2005: 456) are manifold and often stereotypical, but interestingly, such prejudices frequently go unnoticed. Even in an age of political correctness, in which “we don’t freely express judgements about people based on their race or socio-economic status” (Ghomeshi 2010: 9–10), “discrimination on linguistic grounds *is* publicly acceptable” (Milroy and

Milroy 1985: 3). One common prejudice against people committing many orthographic mistakes is that they are stupid or ignorant. Alternatively, non-use of the standard may be attributed to laziness (Ghomeshi 2010: 75), particularly for error types which are flagged by spellcheckers (Figueredo and Varnhagen 2005: 441–445). Since uncorrected errors may thus be attributed to missing attention to detail, perceived laziness infuriates readers more than if mistakes are obviously due to incomplete mastery of the system: as the time devoted to proofreading tends to increase with a text's importance for the author and the status of the addressee, the non-observance of orthographic standards can be interpreted as impolite by addressees who draw conclusions regarding their own status in the eyes of the author.

Furthermore, if a text is inconsistent in its spelling, readers tend to doubt the thoroughness of the author on a more general level, that of content: “if a writer can't get the little things right, he can't be trusted on the big ones” (Booth, Colomb and Williams 2008: 195–196). Even if this may be unjustified, it is in line with the relatively well-working algorithm ‘To assess someone's quality of performance in matters you cannot judge, extrapolate from their performance in matters you understand’. Many people thus seem to generalise from the very basic, superficial level of spelling, which they feel certain about, to other levels that are beyond their evaluative capacity.

Last but not least, correct spelling can be considered a “**marketable asset**” (Ghomeshi 2010: 73). In schools, it is used as a means of educational selection, which ultimately results in social selection (Sebba 2007: 151), and candidates whose CV considerably deviates from spelling norms are less likely to be invited to job interviews, because the mistakes are interpreted as a general sign of carelessness – regardless whether this is correct or not.

However, deviations from linguistic norms do not always achieve the same effect: they are evaluated differently depending on text type and medium, and the effect of their occurrence in a CV differs from that in shopping lists or when they are recognised as purposeful, e.g. as attention-getters in advertising language (Carney 1997: 50). Since receivers of real-time electronic communications and email “are fully aware of the situational constraints under which the message was written” (Crystal 2001: 112), spelling mistakes in these media are “explicitly tolerated, and the pressure to correct them is correspondingly low” (Tavosanis 2007: 101). Furthermore, mistakes only constitute a problem if the readers of a text recognise and interpret them as such – or if the author believes the future readers to do so.

With regard to compounds, even reputable authors may feel insecure about which spelling to use for what compound (Nunberg 1990: 19) but still feel the need to comply with vaguely existing norms, because of the intuition that some compound spellings are perceived as unusual. However, not all language users concede equal importance to the spelling of English compounds in all contexts, and even when spellers consciously hesitate between different variants, some may resort to random guessing without giving the issue more thought, while others might consider reference works. This raises the related question what role is conferred to ‘correct’ spelling in different communities of speakers. Compared to languages like German with relatively rigid spelling conventions, the community of users of English seems to be tolerant regarding spelling variation, but the importance of adherence to spelling norms in education varied in the past decades: a period after the middle of the twentieth century during which spelling was considered unimportant in Great Britain and Australia was followed by the opposed tendency in the 1990s, according to which “achievement in things such as spelling (together with ‘numeracy’)” was “threatening to become all that education is about” (Kress 2000: 11). The current trend at the beginning of the twenty-first century seems to be emphasis on communicative competence, but without completely disregarding formal aspects. This is also reflected in the success of best-selling style guides such as *Eats, Shoots and Leaves* (Truss 2003) and in highly emotional public reactions to orthographic changes such as the omission of hyphens from compounds in the OED (cf. Horobin 2013).

3.2 The Originators of Spelling Norms

The following sections discuss who determines what is ‘correct’ spelling and brings about changes in the orthographic system. In the absence of “officially sanctioned bodies responsible for the regulation of language”, “language usage is often established by a cluster of ‘institutions’, such as lexicographers, grammarians, writers, and editors and publishers” (Modiano 2002: 230). The following sections consider several such institutions, discussing their role in the shaping of English orthography and thus ultimately of English compound spelling.

3.2.1 *Official Institutions*

In contrast to various other languages such as French, Italian, Spanish and Swedish (Modiano 2002: 230), English is not codified by an academy, and

no single official institution is invested with the authority to prescribe the use of English. Note, however, that even the Académie Française, commonly regarded as the prototypical prescriptive institution, is not the 'language police' assumed by common prejudice: according to Vannier (2013), who is involved in the ninth edition of the *Dictionnaire de l'Académie française* (cf. <http://atilf.atilf.fr/academie9.htm>; 18 August 2017), the French academy is not overly conservative (voting unanimously in favour of the latest spelling reform in 1990), it understands its own role as recording persisting usage over several generations and its orthography is merely a recommendation. As a consequence, the frequently claimed opposition between the English language, which has no authoritative institution and is therefore often believed to be chaotic, and French orthography, which is prescribed by its academy (Bollée 1994/1995: 56), is untenable in its strong form.

The fact that there is no academy of the English language can be explained by different factors, such as the death of Queen Anne (who had supported the idea of creating an academy of the English language) in 1714 and the fact that dictionaries and grammars of English were already being written by merchants and booksellers (Tieken-Boon van Ostade 2012: 61, 70). Simon (1980: 12), who supports the idea of founding a new academy of the English language, still believes that this is impossible: English as a global language is spoken not only on the British Isles, but also in the United States, Australia and many more countries in the world, and in contrast to the Icelandic academy (whose standard is unchallenged by significant speech communities using it as a mother language outside the "mother-tongue" nation state"; cf. Modiano 2002: 231) or the Académie Française (which dates from a time at which France was in a position to decide upon the standard of its national language), an academy of the English language would need to consider English in all its varieties. Although learners of larger languages with more than one educational standard often associate prestige with the nation-state in which the language originated, British English has lost its status as the single most influential international variety, and American English has to be regarded as an equally dominant standard at least (Modiano 2002: 233–239), so that no single country or government could now create an academy with binding status for the English language in general.

3.2.2 *The Government and the Educational System*

Spelling can also be determined by official decree within individual countries for individual varieties, as is the case for German (Stang 1993; Augst 2005), whose codified orthography is obligatory for administration and education. Unbeknown to many, there is also an institution in the United States which regulates official government orthography: the US Government Printing Office has been issuing the so-called *GPO Style Manual* since 1984 with its large team of specialists, and 2008 saw the thirtieth edition of that reference work (*GPO Style Manual* 2008: V; for the latest version cf. www.govinfo.gov/content/pkg/GPO-STYLEMANUAL-2016/pdf/GPO-STYLEMANUAL-2016.pdf; 19 August 2017). Nevertheless, its influence is not comparable to that of the German *Duden*, since it does not play the same role in schooling. Great Britain has no similar institution. According to the Cabinet Office, there is no central body and “each department would be required to do their own style manual” (personal communication, 17 May 2010).

The most important way in which governments can influence the spelling of their nations’ language(s) is presumably through education, particularly obligatory schooling. Ministries of Education control teaching through the design of the curriculum and the selection of the reference works to be authorised for teaching at school, with periodic testing on a state-wide scale acting as a control mechanism. Fixed spelling norms are convenient for teachers (cf. Gallmann 2004: 38) because they simplify marking, and the fact that they are imposed from outside conveys an aura of objectivity to orthographic decisions.

However, the curriculum for English in the United Kingdom does not address the spelling of English compounds in much detail: Appendix 1 (Spelling) states that “Compound words are two words joined together. Each part of the longer word is spelt as it would be if it were on its own,” and the only examples given are compounds with solid spelling, e.g. *football* and *playground* (www.gov.uk/government/uploads/system/uploads/attachment_data/file/239784/English_Appendix_1_-_Spelling.pdf, 17 September 2015). At the very end of Appendix 2 (Vocabulary, grammar and punctuation), the content to be introduced in year 6 of English furthermore includes “How hyphens can be used to avoid ambiguity [for example, *man eating shark* versus *man-eating shark*, or *recover* versus *re-cover*]” (www.gov.uk/government/uploads/system/uploads/attachment_data/file/335190/English_Appendix_2__Vocabulary_grammar_and_punctuation.pdf, 17 September 2015). The spelling task mark scheme of the

levels 3–5 English GPS (= grammar, punctuation and spelling) test does not award marks if “a word has been written with the correct sequence of letters but these have been separated into clearly divided components, with or without a dash”, or if “an apostrophe or hyphen has been inserted”, but none of the test words is actually a compound (www.gov.uk/government/uploads/system/uploads/attachment_data/file/327301/2014_KS2_L3-5_English_GPS_mark_scheme_DIGITALHO.pdf, 17 September 2015). All this seems to suggest that the spelling of English compounds does not play a central role in the teaching of English spelling in the United Kingdom at the time of writing. While the acceptable variation for a considerable number of compounds would make marking based on ‘correct’ and ‘incorrect’ spellings difficult compared to other orthographic phenomena taught at school, the raising of pupils’ awareness for degrees of acceptability in compound spelling may seem like a good candidate for future inclusion in the curriculum.

3.2.3 *The Publishing Business*

Orthographic education in the classroom frequently assumes the form of explicit instruction, but spelling is also often learned incidentally from the reading of printed texts. Different originators may be responsible for the form of published texts: the authors (i.e. writers and journalists), the editors and the printers (the last of which once exerted a very strong influence on the spelling of English, which has decreased dramatically; cf. Scragg 1974: 63–74).

Quality writers have traditionally been followed as linguistic models in reference works or style guides: thus Samuel Johnson based his dictionary upon the language of “the best Authors” (Tieken-Boon van Ostade 2012: 63), and McDermott (1990: xv) intends “to survey the current best practice in the use of punctuation by educated writers of English”. While punctuation might not be the focus of interest of the authors themselves, one may assume that good authors are frequently published by publishers with high-quality copy-editing, and that these copy-editors set an example worth following. However, as modern writers tend to experiment with norms (which may involve unusual punctuation as a stylistic device), Hundt (2010: 42–43) assumes that there is no literary canon of present-day English whose texts could serve as an example for language patterns and that the new linguistic models are rather “[w]ell-known politicians, anchormen in television, popular writers of editorial articles in nationwide newspapers” etc. By contrast, publishers’ style guides can be considered “of

fundamental importance in the establishment of current spelling trends” (Scragg 1974: 86) – possibly because in-house style guides are considered authoritative “without any insinuation that their rules are the *only* ones, and the only ones that are good and correct” (Ghomeshi 2010: 74).

3.2.4 *Linguistic Experts*

The group of experts on language is composed of linguists and lexicographers (who are sometimes in personal union). Since the imposition of norms requires authority, linguistic experts should have the advantage of credibility in shaping orthographic norms, but present-day linguistics overwhelmingly considers itself a descriptive discipline, and the majority of linguists presumably only pay attention to punctuation in the context of proofreading texts or marking students’ coursework (Beal 2010: 63). While the general public “view linguists as the seasoned gardeners whose task is precisely to advise on what should be trimmed, removed or promoted in the garden” of language (Burridge 2010: 9), linguists tend to regard spellings such as the greengrocer’s apostrophe as “a matter of proof-reading, not a matter of life and death” or “a symptom of the breakdown of society” (Beal 2010: 63), and are usually reluctant to prescribe or proscribe usage. The public may thus be disappointed by “professional linguists, who can’t be trusted to care about punctuation” (Beal 2012: 187). Linguistic publications which recognise options or linguistic change are regarded as “abrogating their responsibility” (Burridge 2010: 8), since they force their readers to make decisions themselves instead of taking a clear stance. Language users turn to prescriptive reference works in search of black-and-white guidance, because the information they need is easier to access than in descriptive texts, and because the authors’ evaluations make the users’ own choices easier (Busse and Schröder 2010b: 99). As a consequence, the influence of language experts on the norming process has been minimal since the early modern period (cf. Hundt 2010: 37).

To settle compound spelling difficulties, the literature frequently suggests the consultation of a dictionary (e.g. Swan 2005: 551; Reiser 2007), and language users generally seem very willing to accept the information contained in dictionaries as authoritative. For instance, the *Oxford English Dictionary* “has a semi-official standing” (Vallins 1954: 141; cf. also Leisi and Mair 1999: 163). Speakers “tend to believe that the ‘language’ is enshrined” in reference works such as dictionaries and grammars (Milroy and Milroy 1985: 27), and while probably “none but infants” would believe “[t]hat the

weather clerk really makes the weather”, the idea “that language is made by the compilers of dictionaries and grammars is a conception not confined to the young or ignorant” (von Jagemann 1900: 95). The users want their reference works to tell them what is right and what is wrong “because they wish to appear well educated and to eloquently maintain ‘correct usage’” (Allan and Burridge 2006: 120). This results in a naïve trust in the truth of reference works, and those variants included in dictionaries or grammars are unquestioningly accepted as correct (Hanks 1988). This perception stands in stark contrast to the usual aim of present-day lexicographers to record current, generally accepted usage (cf. also Stang 1993: 165), which is derived from corpora of edited language as well as the internet, where spelling is not necessarily checked by copy-editors. If enough language users disagree with an established spelling and use an alternative instead, descriptive reference works are likely to adopt the new variant sooner or later. The observation that dictionaries still differ in their spelling of English compounds (Vallins 1954: 142; Hall 1961: 22; Bauer 2003: 134) can be explained by several reasons, such as corpus-specific variation and the fact that some dictionaries have been updated more recently than others.

3.2.5 *Language Users*

The aforementioned language users represent the last stakeholder in the standardisation process discussed here. Many of these take language matters very seriously, and as late modernity is characterised by “the increased expression of skepticism toward authority and expertise” (Johnson 2002: 567), they consider that their command of the language constitutes the necessary expertise to take an active part in the shaping of normative aspects. Sebba (2007: 133–134) reports several more or less recent attempts at official orthographic reform in German, French, Dutch and Czech, all of which met with considerable public protest. For instance, there was massive resistance against the 1996 German spelling reform, which only affected an average of 0.5 per cent of texts. After the new spelling had been used for some time, several German newspapers returned to the old spelling. The public pressure became so strong that a commission with thirty-seven members was set up by the political authorities of Germany, Austria and Switzerland, and the suggestions proposed by the *Rat für Rechtschreibung* became binding for education and administration in 2006 (Jacobs 2007: 74).

Due to the “intensity of feeling which surrounds orthographic reform”, it tends to meet with “apparently almost inevitable failure or very limited success” (Sebba 2007: 135), particularly in orthographic communities with an important “tradition of literacy, literature, and liturgy”, in which it is “less likely that even minor systematic orthographic change will be freely accepted” and “less likely that any orthographic change will be considered minor” (Fishman 1977: XVI). For the English language, President Roosevelt endorsed 300 spelling changes recommended by the Simplified Spelling Board in 1906 as US government style but was soon overruled (Peters 2004: 511–512), and other attempts at changing English spelling were even less successful (cf. Carney 1994 for a detailed account of failed reform proposals). While most of these reforms were primarily or only concerned with phoneme–grapheme correspondences, some orthographic reforms in other languages have affected the spelling of compounds: since German compounds are traditionally concatenated, there was considerable public protest against the ambiguity in compounds such as *alleinstehend*, ‘single’, whose reformed open spelling *allein stehend* could be interpreted literally as ‘standing alone’ (Jacobs 2007: 71–75). This was changed in the subsequent reform of the reform (cf. earlier in this chapter), and concatenation was recommended for some adjectival compounds involving participles, while others have permitted an additional open variant ever since (Jacobs 2007: 71–75). The Polish orthographic reform of the 1930s also respelled single-word compounds consisting of preposition plus noun as two words, and customers who now had to pay more for their telegrams started quarrels at post offices (Sebba 2007: 148–149). To conclude, attempts at orthographic reform are usually vain if they do not meet with the general approval or at least indifference of the public (cf. also Morton Ball 1939: 39); otherwise they are unlikely to become part of established usage.

Language users seem relatively conservative in language matters (relying on the orthography they took pains to learn as correct at school, and not wishing to give up their knowledge too easily). As a consequence, orthography tends to change only very gradually. Changes originate either from the wish to modify the language or as an unconscious innovation, e.g. because current norms are not known. In doubtful cases, language users either trust reference works or their intuition. To begin with the former, modern reference works are usually written by taking usage into account (cf. earlier in this chapter). Usage is determined based on corpora, and what is ‘usual’ is established by counting (Sepp 2006: 4–5). Since corpora contain texts produced by the language users (some edited, others unedited, some by professional writers, others by non-experts), the language

users may thus influence the next editions of the reference works that they draw on in case of doubt. Intuition as the other potential source of influence on English compound spelling is what the present study intends to approach: investigating the collective gut feeling (cf. Gigerenzer 2007: 16–18) of the users of the English language and determining what makes them select particular spelling variants for particular compounds.

3.3 **Summary**

This chapter presents the normative background of English compound spelling. It discusses underlying concepts such as standardisation, norms and mistakes and explores possible reasons for the observation that language users seem to feel the strong wish to comply with spelling standards – ranging from the requirements of successful communication to considerations regarding power, tradition, status, face, identity, character and economic value. This chapter also discusses the roles of official institutions, the government, the publishing business, linguistic experts and language users as originators of what is perceived as ‘correct’ spelling by the community of users. Only norms which take usage into account will tend to be accepted by the speech community, and codified orthographic norms of English are only binding for particular institutions. English compound spelling is thus not fixed by institutionally sanctioned prescriptive norms, but rather emerges from linguistic use by the community of speakers. Norming involves the mutual influence between language users, who consult dictionaries and write texts, and lexicographers, who read texts and write dictionaries. Non-expert language use is registered by the experts and may be recorded in reference works if usage persists. Since reform requires an authority which has the legal right to change a system, an officially sanctioned reform of English spelling seems impossible in view of the geographical and situational diversification of the English language. As a consequence, small and stepwise orthographic change is more likely to happen.

PART II

Empirical Study of English Compound Spelling

Building on the preceding chapters, which introduced the phenomenon of English compound spelling and delimited the compound concept, the following chapters present a comprehensive empirical study investigating the potential determinants of English compound spelling. The general description of the material and the methods is followed by a detailed discussion of numerous variables that may play some role in spelling variant selection.

Material and Method

The research project described in the following attempts to discern general spelling principles for English compounds, regardless of their part of speech. This sets it apart from previous studies such as that by Sepp (2006), whose exclusive focus on nominal noun+noun compounds allows them to compile lists of potential compounds by applying specified search patterns to tagged corpora while controlling for minimum frequency. Since the present study's wide compound definition recognises a vast amount of possible part-of-speech combinations (cf. 2.5), the extraction of all potential combinations would have resulted in little less than all the words from the corpus sorted in lists of bigrams, trigrams etc. This would have required an immense amount of manual postediting, even after the application of frequency filtering. Furthermore, computer-generated n-grams which do not take into account context may skew the results by ignoring spelling variation due to a compound's occurrence as the constituent of a longer compound (e.g. hyphenated *ice-cream* as the first part of *ice-cream soda* in contrast to its usual open spelling) or in a specific syntactic context (cf. 5.6.2). The list of compounds for the empirical study was therefore compiled from lexicographical material (cf. 4.1) and filtered according to the criteria outlined in Section 2.6. Corpora were then used to determine spelling variation of these Master List compounds in actual usage (cf. 4.2), and a supplementary study investigated corpus-derived compound neologisms that are not listed in dictionaries (cf. Chapter 6).

4.1 Dictionaries

Using lexicographical material (particularly from learner's dictionaries) has several advantages in addition to those mentioned earlier: by drawing on the stock of established words, the analysis starts at the core of the English lexicon and extends into its periphery. This method thus considers only entities which were classified independently as units of the English

language by various linguistic experts – an advantage particularly for researchers with non-English native backgrounds, who might be inclined to regard constructions on the borderline between phrase and compound as compounds if the translation equivalents into their native language are compounds.¹ Furthermore, dictionaries offer the additional advantage of listing lemmatised compound types rather than raw unlemmatised tokens, and as all lexicographers approached for the present study declared to use corpora in their lexicographic work, present-day English dictionaries can be considered a kind of digest corpus search at least to a certain extent.

However, using dictionaries as the starting point has the important disadvantage of posing the danger of circular reasoning: the principles for compound spelling extracted from a sample of reference works may only be representative for established compounds, since these might share certain qualities that make them different from non-listed compounds. One possible way of overcoming this problem would be to restrict all results to established compounds only. The alternative, which was adopted here, is to complement the dictionary-based study by a corpus study (cf. Chapter 6). This procedure permits beginning by describing the core of the compound category, which is of particular importance for the language users, and to test in the next step whether the principles discovered for this subgroup also apply to the more rapidly changing periphery of the compound category.

Table 4.1. lists the dictionaries used in the empirical study. All digital headword lists were generously supplied by the dictionary publishers.² The size of the lists is typically that of a learners' dictionary with 40,000 to 60,000 entries (with the exception of CED, which has more than 130,000 entries). The focus is on British English, which serves as the reference for comparison with American English in the corpus study. All the dictionaries have a similar, relatively recent publication date, which makes them about equally likely to contain particular neologisms and permits a quasi-synchronic consideration of the phenomenon.

¹ Cf. the discussion of the status of English *sandy beach* as a collocation in Herbst (1996, 2011), whose German translation equivalent is a compound and thus concatenated by rule (Duden 2006: 1172–1183).

² Since it was not possible to extract a lemma list with the corresponding part-of-speech codes from Langenscheidt/Collins' *Großwörterbuch Englisch-Deutsch* (2008) nor from the *New English-Irish Dictionary* (NEID; www.focloir.ie) while the latter's compilation was still under way, these dictionaries were only considered in the pilot study but not in the final study reported here. As the publisher's list of compounds from the *Oxford Advanced Learner's Dictionary* contained no solid compounds, a complete lemma list was extracted from the online version of the dictionary at www.oup.com/oald-bin/web_getald7index1a.pl (25 November 2008) with the publisher's consent.

Table 4.1 *Dictionary lemma lists used in the empirical study*

Name of dictionary	Abbreviation	Dictionary type	Lemma type
<i>Longman Dictionary of Contemporary English</i> (5th edn., 2009)	LDOCE	monolingual	list of compounds
<i>Cambridge Advanced Learner's Dictionary</i> (3rd edn., 2008)	CALD	monolingual	complete lemma list
<i>Macmillan English Dictionary for Advanced Learners</i> (2nd edn., 2007)	MED	monolingual	complete lemma list
<i>Taschenwörterbuch Englisch-Deutsch</i> (2007)	TW	bilingual	complete English lemma list
<i>Oxford Advanced Learner's Dictionary</i> (7th edn., 2005)	OALD	monolingual	complete lemma list
<i>Collins English Dictionary</i> (6th edn., 2004)	CED	monolingual	complete lemma list

The majority of the dictionaries are monolingual; in the bilingual dictionary, only the English lemmas were considered.

The part of speech codes of the compounds were standardised across all dictionaries in the following way:

adj	adjective
adv	adverb
conjunction	conjunction
interj	interjection
n	noun
prep	preposition
pron	pronoun
v	verb.

The Master List for the automated dictionary and corpus searches (cf. 4.3) was based on the material from Longman, since the LDOCE lemma list was the most recent headword list, in which all compounds had furthermore been coded by the lexicographers on a descriptive, corpus-based, case-by-case basis (personal communication from Longman/Pearson). To ensure that all Master List compounds corresponded to the present study's criteria for compound delimitation summarised in

Section 2.6, the compounds from the original LDOCE list underwent a manual selection process, and those contradicting the principles were deleted. The LDOCE list items that were not retained include e.g. *gentleman farmer* (which takes the plural on both constituents and is therefore considered a syntactic construction; cf. Donalies 2003: 85) about thirty obscured compounds (such as *starboard* from *steer* + *board*; cf. OED and 2.2.1)³ and a number of affixations such as *do-gooder* (cf. 2.2.2 for the discussion of such borderline cases). For the sake of automatic processing, three LDOCE headwords containing punctuation marks other than blanks or hyphens (namely a slash in *small office/home office* and a comma in *all-singing, all-dancing* and *two up, two down*, respectively) were deleted as well.

The Master List does not contain all the compounds of the English language (which is unfeasible anyway) and not even all the compounds from a particular dictionary (as some lemmas from the printed LDOCE which the present study would have classified as compounds were not included in the original list from the publisher). By contrast, the Master List contains about 10,000 compounds which were classified as such by lexicographic experts and which correspond to the compound principles that the present study subscribes to (cf. 2.6).

Inclusion in many reference works is indicative of a construction's degree of establishment (thus *police community support officer* is only contained in LDOCE and not even in the *British National Corpus*). Since recurrence is a necessary prerequisite for the investigation of variation, most of the present study's analyses focused on compounds which occur at least five times in the six dictionaries with uniquely open, hyphenated or solid spelling for the corresponding part of speech (cf. Chapter 5) and which can be regarded as a relatively central sample within the category of English compounds.

³ The synchronic recognisability of the constituents required by the present study's compound definition (cf. 2.6) is severely reduced when the compound and its constituents differ both in spelling and pronunciation. What can be considered a synchronically recognisable relation needs to be determined on a case-by-case basis. Since compound spelling concerns the written modality, orthographic differences may be assumed to play a more important role in this context than phonetic differences. The presence of orthographic obstacles was therefore used as a filter for compound candidates from the LDOCE list in the empirical study. Compounds with exclusive pronunciation obstacles were retained in this operationalisation to avoid the otherwise necessary systematic check of all pronunciations in external sources. However, the effect of this asymmetrical treatment of spelling and pronunciation on the data is almost negligible: it merely seems to affect the items *breakfast*, *cupboard*, *handkerchief* and *vineyard*, because differences in spelling and pronunciation usually combine (Sanchez 2008: 153).

4.2 Corpora

Table 4.2 provides an overview of all the corpora used in the present study (including those for particular types of text, such as text messages or chat communication). While size is certainly an important criterion in corpus selection, the possibility to carry out diachronic and diatopic comparisons was even more important for parts of the present study. Even if the Brown family corpora only comprise one million words each, the availability of parallel corpora for British and American English in different decades of the twentieth century therefore made them ideal for the analysis of relatively recent language change. Furthermore, Bartsch et al.'s (2015) results suggest that small balanced corpora produce more reliable results than large unbalanced corpora.

A decision in corpus compilation which has a strong impact on the present study is how to treat end-of-line hyphens in the original texts, which are often indistinguishable from compound-internal hyphens. The coding in the Brown corpus with its seventy-character lines usually preserves the conventions of the originals. The words simply run on to the next line; only hyphens at the ends of lines were deleted if they indicate obvious line breaks (e.g. in *situa-tion*), so as to make these words retrievable for corpus searches. Ambiguous end-of-line hyphens potentially realising a compound's hyphenated variant (e.g. in *week-end*) were preserved if the affected words were hyphenated elsewhere in the sample text, but “[i]n all other cases an arbitrary decision to preserve or omit the hyphen was made, based on the general

Table 4.2 *Corpora used in the empirical study*

Corpus	Publication years of content	Variety	Size (in words)
BLOB-1931	1931	BrE	ca. 1,000,000
Brown	1961	AmE	ca. 1,000,000
LOB	1961	BrE	ca. 1,000,000
Frown	1991	AmE	ca. 1,000,000
FLOB	1991	BrE	ca. 1,000,000
BEo6	2003–2008	BrE	ca. 1,000,000
CorTxt	2004–2007	BrE	190,099
NPS Chat Corpus	2006	unclear	ca. 38,500 (10,567 posts)
Blog Authorship Corpus	2004	unclear	> 140,000,000
CompText	2013	BrE	8,864

practice of the text and the listings in Webster's *New International Dictionary, Third edition*" (Francis and Kučera 1979). Since these decisions were all recorded in the manual, it was possible to automatically search and manually extract all the instances of arbitrary hyphens or non-hyphens in the Brown corpus. One hundred sixty-two word types (of which six are repeated) were affected in total; the majority of them compounds, but also some instances of what the present study classifies as prefixations (*overage*), suffixations (*teenager*) or neoclassical compounding (*neo-dadaist*). The compilers' decisions for a hyphen (126 cases) very clearly outweigh those against a hyphen (37 cases), with the compound type *long+time* occurring in both categories. To determine how many of these items also occur in the LDOCE Master List, they were formatted according to the conventions of the present study and plural forms were manually complemented by their singular. The automatic *Excel*-based comparison with the Master List yielded an overlap of only sixty compound types, so that the influence of this factor should not be overestimated.

Concerning hyphenation, Johansson, Leech and Goodluck (1978) state in the corpus manual of the Lancaster-Oslo/Bergen (= LOB) Corpus that:

14.3 A line-end hyphen in the source text is not coded, except where the hyphen is part of the normal spelling of the word.

14.4 Where spelling practice varies with regard to hyphenation, a coding decision has to be made as to whether the line-end hyphen is preserved in the coded text or not. The hyphen is preserved:

(a) if dictionaries show that hyphenation is normal.

(b) if the word in question is hyphenated elsewhere in the same text.

14.5 In other cases, where doubt still remains, the line-end hyphen is included or excluded according to the judgment of the coder.

The approach is similar to that of the Brown Corpus, although slightly less systematic (as it may involve the coder's judgment) and less well documented: while dictionaries may differ in their treatment, the LOB corpus manual does not mention which reference works were used and what was done in conflicting cases. Nor does it list the compounds for which coding decisions had to be taken, and only among the instances of 'variant spelling' (which includes e.g. *-ise* vs. *-ize*) does it mention eight instances of differing compound spelling within the same text (e.g. *lace-making* vs. *lacemaking*).

By analogy to the original Brown corpus – and in contrast to LOB – the user manual of *Freiburg-LOB* (= FLOB; Hundt, Sand and Siemund 1998) also contains an indication of what are called 'ambiguous hyphens', marked with the code `<?_>-<?/>`. Following the procedure described

earlier, the affected words were extracted, post-edited and compared with the Master List, yielding an overlap of ninety compounds.

The index files of the Frown corpus (Hundt, Sand and Skandera 1999) relate the reader to the FLOB manual for information about sampling and mark-up conventions. One may therefore assume that hyphens were treated in the same way as in FLOB – and indeed, the code `<?_>-<?/>` used in FLOB for ambiguous hyphens is among the mark-up codes for Frown. The analysis of the data along the same lines as those outlined earlier yielded an overlap of 166 compound types with the Master List.

As for the more recent members of the Brown corpus family, at the time when the present study was conducted, BLOB-1901 (which stands for ‘before LOB’; <http://ling.lancs.ac.uk/profiles/Geoffrey-Leech/>, 05 April 2011) was still under construction, but BLOB-1931 had already been compiled. Its texts, which date from 1931 +/- a period of three years (www.helsinki.fi/varieng/CoRD/corpora/BLOB-1931/index.html, 18 August 2017), were usually scanned and then extracted from the program Omnipage OCR, with a minority of texts typed in manually using *Word* (here and in the following: personal communication from Nicholas Smith). The OCR proofreader’s default hyphenations were accepted, unless they were obviously not in line with contemporaneous norms in the same text or elsewhere in the corpus – but that was hardly ever the case, except for rare items such as time adverbs (*to-day*, *to-morrow*) or street names (*Bunhill-row*, *Brewery-road*). Occasionally, the code `&rehy;` was inserted to indicate ambiguous hyphenation. By analogy to the other corpora, all such cases were extracted and compared with the present study’s Master List, yielding forty-six relevant compound types.

The majority of the texts in the BEO6 corpus, which extends the Brown family into the present, were published between 2005 and 2007, with the year 2006 as the median sampling point (www.ling.lancs.ac.uk/profiles/Paul-Baker/, 01 April 2011). Although the publication in paper form was a requirement for inclusion in the corpus, all texts were collected from online sources before or after their print publication (www.helsinki.fi/varieng/CoRD/corpora/BEO6/index.html, 18 August 2017). The texts were copied and pasted into *Word* documents and saved as text only, and end-of-line hyphenation was preserved like in the originals and not marked in any way (personal communication from Paul Baker). Since HTML texts do not usually contain end-of-line hyphenation (www.mnn.ch/hyph/hyphenation1.html, 26 October 2013), there were presumably no end-of-line

hyphens in the corpus texts created from this kind of format, but there might have been some end-of-line hyphens in the corpus texts derived from pdfs. Since it was not possible to obtain a text version of BEO6 for copyright reasons (in contrast to all the other corpora listed earlier), the information on the spelling of the compounds had to be extracted automatically from the web interface in CQPweb (<http://cqpweb.lancs.ac.uk/be2006>, 30 May 2011). The compounds were converted into strings with the three spelling variants and their possible internal variations (e.g. with regard to plural; cf. Section 4.3). These were subsequently built into a URL, which was sent to CQPweb. The frequency was then extracted from the return value.

The next three corpora listed in Table 4.2 are not part of the Brown family and represent specific communicative situations rather than the English language as a whole. They were used to control for spelling differences in unedited vs. edited data and made it possible to consider the influence of temporal and spatial constraints on English compound spelling.

Release 1.0 of the NPS Chat Corpus (Forsyth and Martell 2007) contains 10,567 chatroom posts gathered in 2006 (<http://faculty.nps.edu/cmartell/NPSChat.htm>, 18 August 2017), which were not coded for the authors' geographical origin. In the original dataset, each line begins with the user name followed by a colon and a space, but this information was deleted for the present study. End-of-line hyphenation did not have to be adjusted for, since chatroom contributions simply switch to the next line if a word is too long.

The Blog Authorship Corpus (Schler et al. 2006) assembles the posts gathered from the internet site www.blogger.com in August 2004. It contains more than 140 million words in 681,288 posts comprising at least 200 "common English words" by 19,320 bloggers. There are equal numbers of male and female bloggers in each age group (13–17, 23–27 and 33–47; <http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm>, 18 August 2017). The separate html files downloaded from the corpus site (which contain no end-of-line hyphens) had to be converted into a single text file for the empirical study. The sociological data provided by the bloggers themselves (gender, age, industry and astrological sign), which were coded in the original filenames, were lost in that process.

The CorTxt Corpus (Tagg 2009) assembles 11,036 text messages comprising a total of 190,099 words. In the corpus specifications provided with the corpus, Tagg states that most of these text messages were collected from the researcher's friends and family between March 2004 and May 2007 and

complemented by 441 messages from an anonymous public AOL forum for forwarding text messages. The majority (but not necessarily all) of the authors can therefore be assumed to be native speakers/writers of British English (although there is a known proportion of 4 per cent non-native English contributions).

In contrast to all of the previous corpora, the CompText corpus was compiled specifically for the present study in order to test the spelling algorithms derived from the data and to determine whether the features established for the compounds with clear spelling preferences can be used to predict the spelling of less established compounds (cf. Chapter 6). The CompText corpus contains 8,864 words from twelve British English texts belonging to different registers (newspaper articles, academic articles, nonfiction books and miscellaneous texts – a glossy article, a recipe and a sports article), with each text comprising between 385 and 1,028 words. Texts were considered British if they were published by a British publisher or appeared in a Britain-based (online) newspaper or journal and/or if the author was known to be of British origin, e.g. for the academic articles and the books. All texts are included from the very beginning (including the headline and information about the authors of articles – but excluding photograph captions or links), with the cut-off point at the end of the sentence exceeding the 1,000-word mark in texts with more than 1,000 words. The selection of the texts was randomised as much as possible, e.g. by choosing the articles based on their prominence as the top stories of online newspapers or the recipe of the day. The texts were not read prior to their inclusion in the corpus in order to avoid any bias in the selection. End-of-line hyphenation did not result in any doubtful cases.

To conclude, the present study did not modify the texts of the corpora discussed earlier except when deleting or modifying mark-up codes (e.g. word-internal diacritics, such as *e-acute*), so that these would no longer interfere with the search for compounds. The deletion of codes in angled brackets was equivalent to accepting all ambiguous hyphens codified as `<?_>-<?/>` in Frown and FLOB. The following variation between the corpora in their treatment of end-of-line hyphenation was considered for the automated corpus searches (cf. 4.3):⁴

⁴ In contrast to the usual convention also followed by the other corpora, dashes in Brown and LOB are not surrounded by blanks but directly connected to the preceding word (e.g. `<text- text text text-text>`). While this might potentially lead to confusion with hyphenated compounds at line ends, the interference with actual compounds is presumably minimal.

Table 4.3 *Make-up of the CompText corpus*

	Text	Register	Source/Author	Date of publication	Length (in orthographic words separated by spaces)	Number of compounds (tokens)
1	“David Cameron rebukes ministers for saying they would vote to leave EU”	Newspaper article (online)	<i>The Guardian</i>	13 May 2013	774	47
2	“Tia Sharp murder: Stuart Hazell changes his plea to guilty”	Newspaper article (online)	<i>The Telegraph</i>	13 May 2013	590	33
3	“Revealed: Eerie new images show forgotten French apartment that was abandoned at the outbreak of World War II and left untouched for 70 years”	Newspaper article (online)	<i>The Independent</i>	13 May 2013	385	28
4	“Chris Huhne and Vicky Pryce released early from prison”	Newspaper article (online)	<i>The Sun</i>	13 May 2013	507	49
5	“Materials against materiality”	Academic article	Tim Ingold, <i>Archaeological Dialogues</i>	2007	1,028	40
6	“The uses of value”	Academic article	Daniel Miller, <i>Geoforum</i>	2008	1,002	57

7	<i>One Man and His Bike</i>	Nonfiction book	Mike Carter	2011	1,017	44
8	<i>Empire</i>	Nonfiction book	Jeremy Paxman	2012	1,012	49
9	<i>Atlantic: A Vast Ocean of a Million Stories</i>	Nonfiction book	Simon Winchester	2010	1,013	76
10	“Behind the scenes at the TV BAFTAs: Celeb beauty secrets”	Women’s magazine article (online)	<i>Cosmopolitan UK</i>	13 May 2013	586	72
11	“Arsenal 4 Wigan 1”	Sports article (online)	BBC Online	14 May 2013	542	52
12	Mexican street salad	Recipe (online)	Jamie Oliver	13 May 2013	408	35
	TOTAL				8,864	582

- BLOB-1931, BEO6, CorTxt, NPS Chat Corpus and Blog Authorship Corpus represent the unmarked standard case with lines running on until they end with a paragraph break. Compounds in these corpora are concatenated or contain an internal hyphen or space.
- The length of lines in LOB and Brown, by contrast, is restricted: the insertion of hard returns (¶) in the compilation of those corpora may have resulted in paragraph breaks in compound-medial position, e.g. in *apple¶pie*. As a consequence, the paragraph may take on the value of a compound-medial space in these corpora.
- FLOB and FROWN also contain paragraph breaks. While these were not generally counted as corresponding to spaces in their function, the two corpora differ from the others in their occasional use of a paragraph break followed by a space, e.g. in *apple¶ pie*. Only this combination of paragraph break and space was considered to count towards open spelling.

To sum up, even though end-of-line hyphenation is not treated absolutely identically in all of the corpora mentioned earlier, they are sufficiently similar in their structure and treatment of the phenomenon to permit a comparison. While the results may be slightly skewed towards hyphenation due to the occurrence of ambiguous hyphens at line ends, the documentation of the corpora at least permits the determination of the relatively small number of cases affected. Furthermore, no other set of corpora is so similar in its structure and permits a comparison between both regional and diachronic varieties of English (the ICE corpora, for instance, only permit the comparison of geographical differences; cf. <http://ice-corpora.net/ice>, 18 August 2017).

4.3 CompSpell

The PHP program CompSpell was written specifically for the present study to carry out automatic searches for all orthographic variants of the Master List compounds in the dictionary-based word lists and corpora. While it was possible to use spaces, hyphens and cell boundaries for the automatic recognition of constituent boundaries of hyphenated and open compounds, the cut-off points of the constituents in solid compounds had to be coded manually by inserting a plus sign (+) in the Master List words (e.g. *land+mine*). This was necessary because the check against a list of possible constituents might have resulted in nonsensical analyses such as *pup+pet*, *rat+tan* or *ton+sure*.

Manual segmentation was restricted to the highest level of word formation. Since the first part of compounds such as *nineteenth hole* is a compound which has then undergone suffixation – i.e. [*nine+teen*] +*th* – it was left unchanged, so that the combination of *nine* and *teen* remained invisible to the program. Eleven compounds contained a constituent with a prefix set off by a hyphen or a hyphenated reduplicative, e.g. *anti-virus software* and *yo-yo dieting*. Since these hyphens may vary in different reference works even though they are not the types of hyphen considered in the present study, the following compounds were deleted from the Master List:

air vice-marshal
anti-lock braking system
anti-virus software
auto-immune disease
extra-sensory perception
hand-eye co-ordination
non-commissioned officer
non-executive director
post-traumatic stress disorder
semi-skimmed milk
yo-yo dieting.

CompSpell counted all spelling-sensitive occurrences of the Master List compounds combined with their part of speech in the dictionaries CALD, CED, LDOCE, MED, OALD and TW. For example, the adjectival compound *home+grown* occurs six times in the dictionaries, namely zero times with open spelling (O), five times with hyphenated spelling (H) and once with solid spelling (S). The resulting database contains separate columns for each possible combination of O, H and S with up to three constituents (e.g. OH, OS) and one column for the fifty-four Master List compounds with four or more constituents. The program searched for the complete content of the cells in the dictionary word lists to avoid accidental hits for the target compound used as the constituent of a longer compound. This prevents finding e.g. the hyphenated form *ice-cream* as a potential spelling of the target compound *ice+cream*, in spite of the fact that all dictionaries actually spell *ice+cream* open (in contrast to its hyphenated use as the first part of *ice-cream soda*, which would have produced the unwanted hit if no measures had been taken).

A number of automatic tolerance principles were applied by CompSpell concerning the use of punctuation, capitalisation, diacritics and inflection.

In the corpora, like in the dictionary lemma lists, CompSpell searched for the open, hyphenated and solid spellings of the compounds from the Master List and computed the respective frequencies of the three orthographic variants (O, H and S). CompSpell ignored punctuation marks before or after the whole compound (such as quotation marks, brackets or commas), but if sequences with open spelling were interrupted by punctuation marks such as a comma or a full stop, these sequences were not included in the compound count, and only the potential spelling variants <text text>, <text-text> and <texttext> were retrieved by CompSpell (cf. 4.2 for the corpus-specific variation in line breaks considered by the program). To allow for capitalisation in the beginning of sentences, titles etc., the search was made case-insensitive. Diacritics (e.g. the acute accent in *attaché case* or *matinée jacket*) were also ignored.

To retrieve a maximum of hits from the comparatively small corpora used and to find inflected compounds, CompSpell tolerated the addition of particular strings of characters to the right of the target compounds (cf. later in this section). The CompSpell tolerance principles, which draw on Swan (2005: 113–115, 394, 514, 551–554), Ungerer et al. (1984: 95, 101) and Quirk et al. (1985: 97–105), were only applied to compounds with the corresponding part of speech, so that the search for a noun compound like *hand+hold* returned nominal forms like *hand+hold's* but no inflected verbal forms like *hand+held*. The addition of inflection was only admitted at the end of the compounds but not in the middle, since the proportion of compounds in English that add their plural morpheme inside the compound (e.g. *mothers-in-law*; cf. Quirk et al. 1985: 313) is very small. Note that the genitive singular (*palm tree's*) and the genitive plural (*palm trees'*) are covered by CompSpell's general acceptance of punctuation marks directly preceding or following the compounds and their variants. The CompSpell principles were checked against the regular expressions of Thomas Proisl's lemmatiser, which is based on the LEMMINGS lemmatisation rules from the BNC Sara server. Where no compounded example words could be found in the present study's material, the corresponding rules were still included for the sake of completeness, so as to permit their application to other datasets.

Since adverbs are basically inflected like adjectives (Ungerer et al. 1984: 101), identical tolerance rules are applied to compounds of either part of speech. English pronouns share certain characteristics with nouns, but while many pronouns have a genitive case (e.g. *someone's*), number in pronouns is usually morphologically unrelated (e.g. singular *he* vs. plural *they*) except in *yourself* vs. *yourselves* (Quirk et al. 1985: 336, 343), which is

Table 4.4 *CompSpell inflection tolerance principles (= part-of-speech-specific inflection which may be added to the compounds in the corpus searches)*

n:	Ø	<i>palm tree</i>
	s	<i>palm trees</i>
	sh#/ch#/s#/	<i>council taxes</i>
	x#/z# +es	(= If a noun compound ends in <sh> etc., CompSpell accepts the addition of final <es>.)
	C _y # + ies	<i>cocktail parties</i> (= If a noun compound ends in a consonant followed by <y>, CompSpell accepts the substitution of the final <y> with <ies>.)
v:	Ø	<i>spoon-feed</i>
	s	<i>spoon-feeds</i>
	es	<i>second-guesses</i>
	C _y # + ies	<i>blow-dries</i> (= If a compound verb ends in a consonant followed by <y>, CompSpell accepts the substitution of the final <y> with <ies>.)
	ing	<i>spoon-feeding</i>
	e# +ing	<i>touch-typeing</i> (= If a compound verb ends in <e>, CompSpell accepts the substitution of the final <e> with <ing>.)
	VC# +Cing	<i>sidestepping</i> (= If a compound verb ends in a single consonant following a single vowel, CompSpell accepts the repetition of that consonant followed by <ing>.)
	c# +king	<i>panicking</i> (= If a compound verb ends in <c>, CompSpell accepts the addition of final <king>.)
	d	<i>chain-smoked</i>
	ed	<i>brainwashed</i>
	C _y # + ied	<i>blow-dried</i> (= If a compound verb ends in a consonant followed by <y>, CompSpell accepts the substitution of the final <y> with <ied>.)
	y# + id	<i>gainsaid</i> (= If a compound verb ends in <y>, CompSpell accepts the substitution of the final <y> with <id>.)
	VC# +Ced	<i>pistol-whipped</i> (= If a compound verb ends in a single consonant following a single vowel, CompSpell accepts the repetition of that consonant followed by <ed>.)
	c# +ked	<i>panicked</i> (= If a compound verb ends in <c>, CompSpell accepts the addition of final <ked>.)
	n	<i>known</i>

Table 4.4 (*cont.*)

	en	<i>fallen</i>
	ie # + ying	<i>zip-ieying</i> (= If a compound verb ends in <ie>, CompSpell accepts the substitution of the final <ie> with <ying>.)
adj/adv:	Ø	<i>fast</i>
	er	<i>older</i>
	est	<i>oldest</i>
	r	<i>later</i>
	st	<i>latest</i>
	Cy# +ier	<i>happy-go-luckyier</i> (= If a compound adjective/adverb ends in a consonant followed by <y>, CompSpell accepts the substitution of the final <y> with <ier>.)
	Cy# +iest	<i>happy-go-luckyiest</i> (= If a compound adjective/adverb ends in a consonant followed by <y>, CompSpell accepts the substitution of the final <y> with <iest>.)
	VC# +Cer	<i>fatter</i> (= If a compound adjective/adverb ends in a single consonant following a single vowel, CompSpell accepts the repetition of that consonant followed by <er>.)
	VC# +Cest	<i>fattest</i> (= If a compound adjective/adverb ends in a single consonant following a single vowel, CompSpell accepts the repetition of that consonant followed by <est>.)

not considered a compound but a suffixation in the present study. This made it possible to omit the category of inflectional endings for pronouns in CompSpell altogether, since genitives (e.g. *each other's*) were already covered by punctuation rules.

While one may expect a certain proportion of noun and verb compounds to take inflection, the majority of adjectives and adverbs can be expected to occur in the base form: only few adjectives and adverbs with two syllables take a synthetic comparative and superlative, since the comparison with *more* and *most* is more common for longer members of these parts of speech (Swan 2005: 114). Adverbs with three syllables are almost exclusively analytic (Swan 2005: 114), and those ending in *-ly* always form the comparative and superlative with *more* and *most* according to Ungerer et al. (1984: 101). Note, however, that the comparative and superlative forms *happy-go-luckier* and *happy-go-luckiest* were attested with 806 and 12,700 Google hits

in July 2017, which might suggest that the number of syllables determining the choice between the analytic and synthetic comparatives and superlatives for compound adjectives might be based only on the constituent to which the suffix would be attached.

While the application of these tolerance principles theoretically includes agrammatical forms (e.g. the incorrect past participle **known*, or double plurals such as **bicycle shortses* for compounds which are already pluralised in the Master List), such hypothetical forms have no negative effect on the results, because they simply yield no hits. Consequently, the degree of precision of the tolerance principles was deemed sufficient for the present purpose.

In order to permit the retrieval of irregularly inflected forms of the Master List compounds in corpora (e.g. *fieldmice* as the plural of *field-mouse*), a list of irregular forms corresponding to the various parts of speech was compiled by inverting a list of irregular forms computed from the English dictionary of the Unitex corpus processing system (<http://info.lingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html>, 31 May 2011) to yield all inflectional forms for each lemma. All forms covered by the regular principles outlined earlier were deleted, and the following items from Swan (2005), Ungerer et al. (1984) and Quirk et al. (1985) were added manually to the exception list considered by CompSpell:

woman	women
be	am/are/is/was/were/been
have	has/had
do	did/done
well	better/best
badly	worse/worst
late	latter/last
far	farther/further/farthest/furthest
near	next
little	less/lesser/least
much	many/more/most.

The exception lists apply in addition to the base forms and the rules, so that e.g. the search for the word *near* would find all corpus items

- with the base form *near*
- with the rule-based forms *nearer* and *nearest*
- with the irregular form *next*.

CompSpell also considers yet another potential type of variation within compound constituents, namely alternative spellings such as *color/colour*, most of which appear to be linked to the differences between British and American English. Since the *Longman Dictionary of Contemporary English*, which the Master List is based on, comes from a British publisher, the main entries – and thus the Master List words – usually adopt British English spelling. The American variants usually take the form of a cross-reference (e.g. at *color*: “the American spelling of COLOUR”), but this rarely applies to compounds (e.g. from *sulfur dioxide* to *sulphur dioxide*). In order to permit the retrieval of compounds with both British and American English spelling from the corpora (e.g. *colour+coded* and *color+coded*), CompSpell uses Roland Grant’s comprehensive list of more than 1,700 roots and derivatives with British and American English spelling variants (www.tysto.com/articles05/q1/20050324uk-us.shtml#note, 27 April 2011), but the effect of this group of compounds need not be overestimated.

While it was attempted to retrieve the Master List compounds in the corpora as completely as possible, particular compounds may have escaped CompSpell’s searches in certain contexts, e.g.:

- in elliptical compounds with a parallel structure (*They were selling both **carriage and grandfather clocks***). Since the compound *carriage clock(s)* is interrupted by the insertion of the sequence *and grandfather*, the program cannot find it.
- in the presumably extremely rare instances where compounds are interrupted by punctuation marks, e.g. brackets indicating potential omissibility such as (*adventure*) *playground*.
- in cases of incorrect coding in the corpora, e.g. due to missing spaces between a compound and neighbouring constructions (e.g. **She was waiting at the**bus stop***).
- in cases where the word from the Master List is already in the plural (cf. earlier in this chapter). While the majority of cases seems to forbid a singular, there are exceptions, such as *air brakes*, whose existing singular *air brake* would not have been found.

In addition, the following few factors might slightly skew the results:

- The number of hits with open spelling will tend to be too high, since some open compounds are formally identical with phrases (e.g. *a green house* meaning ‘a house that is coloured in green’ instead of the intended *greenhouse* ‘glass building’). This could be the case where the

part-of-speech combinations within a compound are also possible on the syntactic level, e.g.

adj + n	<i>soft+ware</i>	(n)
num + adj	<i>80+odd</i>	(adj)
v + particle ⁵	<i>see+through</i>	(adj).

- The number of hits with hyphenated or solid spelling may be slightly de- or increased, depending on how hyphens at the ends of lines – which could either be true hyphens or end-of-line hyphens – are treated in the corpora (cf. 4.2). The number of cases affected in the present study is, however, relatively small.
- Hits for compounds which are contained within longer compounds will count towards the frequency of both list words. Thus the corpus search for *peanut* will also include the instances that are found for the compound *peanut butter*.

If one considers all of this, the conclusion to draw is that corpus studies which do not take spelling variation into account will miss a more or less important portion of the data, depending on their research question. The larger the corpora, the less important the influence of this factor – but in studies based on small corpora such as those from the Brown corpus family, it makes sense to apply particular care.

⁵ Following Herbst and Schüller (2008: 61–67), *particle* is understood here as a cover term comprising traditional prepositions, traditional subordinating conjunctions and certain types of traditional adverbs.

Potential Determinants of English Compound Spelling

Considering how commonly English compound spelling is believed to be completely idiosyncratic, one may assume that any principles underlying it (if there are any) are either too complex to be immediately recognisable or consist of the interaction of a large number of factors, or both. The following sections therefore investigate an extensive number of variables which might potentially exert some influence on variant selection in the spelling of English compounds. These variables concern different areas of language, namely spelling (5.1), length (5.2), frequency (5.3), phonology (5.4), morphology (5.5), grammar (5.6), semantics (5.7), diachronic aspects like age of the compound (5.8), discourse aspects like regional variation (5.9), systemic aspects like analogy (5.10) and also some extralinguistic aspects like economy (5.11). Eventually, the specific variables (marked with square brackets such as [PoS_comp] for ‘part of speech of the compound’) are combined into super-variables such as heterogeneity or complexity of the constituents (5.12). In order to increase readability and to avoid repetition, the full empirical analysis for the investigated variables is given in subsections introducing the hypothesis, method, results and discussion. Summaries at the end of each larger block give an overview of the most important results.

In contrast to previous research, which has mainly focused on variation in compound spelling (cf. 1.1.1), the present study is based on the assumption that the principles derived from prototypical compounds with identical spelling in many dictionaries can be used to determine the most likely spelling(s) of other compounds. This hypothesis was tested by analysing 600 core compounds, the OHS_600 sample, in great detail. The OHS_600 list combines three subsets of the Master List which comprise 200 exclusively open, 200 exclusively hyphenated and 200 exclusively solid biconstituent compounds occurring in at least five of the six dictionaries used (cf. 4.1) and selected at random from the compounds fulfilling these requirements. A larger sample of biconstituent compounds with

exclusively open, hyphenated and solid spellings occurring in at least five dictionaries (OHS_extra) was sometimes drawn upon in order to meet the requirements for statistical significance testing for variables with infrequent values. These 3,864 compounds, for which only automatic and semi-automatic variables were coded (cf. 5.13), do not overlap with the OHS_600 sample.¹ Two more subsets of the Master List compounds permitted specific testing: the Master_5+ sample contains the 1,196 compounds occurring at least five times in the dictionaries with the exclusion of the OHS_600 and OHS_extra items. Its compounds either comprise more than two constituents or vary in their spelling to some extent. Finally, the Master_1–4 sample contains the 4,381 compounds from the Master List which conform to the present study's compound criteria but occur in only up to four dictionaries.

SPSS was used for most of the statistical testing. Since more than sixty hypotheses were tested, the following sections reduce lengthy repetitions in order to be more reader-friendly. Technically, it is always the null hypothesis which is tested – if it is true, the alternative hypothesis has to be rejected; if it is not true, the evidence is regarded as favouring the alternative hypothesis (cf. Moore and Notz 2006: 452–453) – but the analyses frequently skip this intermediate step on the surface while maintaining it implicitly. Furthermore, the discussion of the data from tables with high statistical significance sometimes goes beyond the analysis of the hypothesis tested. Since a detailed formal analysis would have involved an additional alternative hypothesis with a corresponding null hypothesis, a shorthand form was also used in the discussion of such data for the sake of readability.

5.1 **Spelling**

While it may initially appear tautological to consider spelling as a determinant of spelling variant selection, certain aspects regarding the spelling of the constituents may indeed play a role in determining whether particular English compounds tend towards open, hyphenated or solid spelling.

5.1.1 *Graphotactics*

The most important graphic aspect that may influence the spelling of English compounds concerns their graphotactics. Following Neijt (2002),

¹ Tests carried out on OHS_extra for a number of hypotheses generally supported the OHS_600 results (except for Hypotheses B5 and C3).

graphotactics is understood here as the equivalent of phonotactics in the written modality, i.e. as a description of the principles determining what are permissible combinations of graphemes in a language. Since most graphemes are linked to a phonetic representation (even if there is no one-to-one correspondence; cf. Venezky 1967), some graphotactic principles correlate with phonotactics, and one may expect spelling to reflect syllable structure at least to a certain extent.

5.1.1.1 *Hypothesis A1 – Four or More Consonant Graphemes across the Constituent Joint*

The occurrence of four or more consonant graphemes across the constituent joint (e.g. in *all+star*) disfavours solid spelling. → Refuted.

Since English syllables are built around a vocalic core (unless they contain a syllabic consonant such as /n/, /m/, /ŋ/, /l/ or /r/; cf. Gut 2009: 76), written sequences of letters can be expected to contain a vowel grapheme at least after a certain number of consonant graphemes. Kiraz and Möbius (1998: 71) state that a maximum of three consonant phonemes is used in the onset and a maximum of four consonant phonemes in the coda of present-day spoken English syllables. The permissible number of consecutive consonant graphemes must be even higher, because some graphemes (e.g. in *numb*) do not correspond to a phoneme, and because some grapheme combinations correspond to single phonemes whose articulation is unrelated to the phonetic default realisation of the individual graphemes (e.g. <ph> => /f/, as in *photo*; cf. Venezky 1967: 85). If many consonant graphemes at the end of the last syllable of one constituent and at the beginning of the first syllable of the next constituent are joined in a new compound, the result is an extremely long sequence of consonant graphemes. As a consequence, language users may tend to use either the open or the hyphenated compound spelling variant in order to make segmentation easier (cf. also 1.1.1). This can be formulated as Hypothesis A1:

A1: The occurrence of four or more consonant graphemes across the constituent joint (e.g. in *all+star* or *ash+tray*) disfavours solid spelling.²

² The corresponding H₀ would be “The occurrence of four or more consonant graphemes across the constituent joint does not disfavour solid spelling”, but a shorthand notation will be used in the following for ease of reading. *Disfavour* is used to refer to negative correlation; *favour* to positive correlation.

The number four was chosen as the cut-off point, since this involves either a minimal two-consonant cluster both at the end and at the beginning of adjacent constituents or one slightly longer three-consonant cluster followed by a single consonant. In order to test Hypothesis A1, the program CompSpell computed the number of consonant graphemes at the constituent joint, provided that there were consonants on both sides of the joint. Otherwise, the consonants were ignored, and *x* was entered into the respective column. Since the grapheme <y> may stand either for a vowel or for a consonant in the pronunciation, it had to be determined how to treat this grapheme with regard to consonant clusters: <y> was considered a vowel if it preceded a consonant (*byte*) or a word boundary (*prey*). If <y> was followed by a vowel, it was considered a consonant (*you*), except if it was preceded by another vowel (*eye*); in this latter case, it was counted among the vowels.

Pearson's chi-square test was conducted in SPSS for the list of the 600 OHS_600 compounds with the dependent variable 'compound spelling' [OHS] and the independent variable 'number of consonants occurring across the joint between the first and second constituents' [o_1_CC_2]. Since 33.3 per cent of the cells contained expected counts below five, and since three expected counts were smaller than one, there was not enough data to permit a statistical test of the hypothesis in this detailed format (Moore and Notz 2006: 496). The data were therefore recoded in such a way that four or more consonants in a row were treated jointly [o_1_CC_2_r],³ and as a cluster was considered as consisting of at least two consecutive consonants; zero or one consonant at the boundary were also treated jointly.

Grouping the higher values raised the level of significance to $p = 0.020$, but while we can conclude that the number of clustered consonants seems to play a small role in English compound spelling (in contrast to Kuperman and Bertram 2013: 950, whose study found no effect of consonant clusters across constituent boundaries on spelling choice), Hypothesis A1 still needs to be refuted: four or more consecutive consonant graphemes across constituent boundaries do not disfavour solid spelling – as is exemplified by the compounds *birthplace* and *matchstick*, which contain sequences of five consonant graphemes at their constituent boundaries. The expected number of 19.7 solid compounds with four or more consonants in a row was even exceeded by the actual number of 24 instances.

³ CompSpell incorrectly classified *cat's cradle* as containing six consecutive consonants.

Table 5.1 *Grouped consonant clusters across the constituent joint [o_1_CC_2_r] and spelling in OHS_600*

			Number of consonants across constituent joint				Total
			0-1	2	3	4-6	
OHS	o	Count	82	50	57	11	200
		Expected Count	74.7	51.0	54.7	19.7	200.0
		% within o_I_CC_2_r	36.6%	32.7%	34.8%	18.6%	33.3%
	h	Count	84	43	49	24	200
		Expected Count	74.7	51.0	54.7	19.7	200.0
		% within o_I_CC_2_r	37.5%	28.1%	29.9%	40.7%	33.3%
	s	Count	58	60	58	24	200
		Expected Count	74.7	51.0	54.7	19.7	200.0
		% within o_I_CC_2_r	25.9%	39.2%	35.4%	40.7%	33.3%
	Total	Count	224	153	164	59	600
		Expected Count	224.0	153.0	164.0	59.0	600.0
		% within o_I_CC_2_r	100.0%	100.0%	100.0%	100.0%	100.0%

This finding was also replicated for the OHS_extra compounds after grouping the data to permit statistical testing. While the chi-square test for the dependent variable ‘compound spelling’ and the grouped independent variable [o_1_CC_2_r] resulted in a significant effect (p = 0.000), there was no significant reduction of solid spellings in compounds with long consonant clusters across the constituent joint (with 126 instances as against the expected 129).

5.1.1.2 *Hypothesis A2 – Identical Graphemes across the Constituent Joint*

The occurrence of identical graphemes before and after the constituent joint – e.g. in *felt+tip* – disfavours solid spelling. → Confirmed.

According to Rohdenburg’s (2003: 278) principle of *horror aequi* (‘the fear of the same’), language users avoid the adjacent repetition of identical entities or structures in grammar, and Mondorf (2003: 278) points out similar effects in phonology and morphology. In compounding, grapheme doubling or even tripling is not avoided, but its visual impact may be attenuated by separating identical consonants or vowels by a space or hyphen. This is in line with the advice of some style guides to use hyphenation (and sometimes also open spelling) to avoid the confusion caused by double or triple graphemes at constituent

Table 5.2 *Identical graphemes [Ident_lett]
across the constituent joint in OHS_600*

Double grapheme	Frequency
<i>none</i>	585
c	2
e	4
h	3
l	1
m	1
r	3
s	1
t	2
w	2

boundaries (e.g. Morton Ball 1951: 3). Hypothesis A2 consequently summarises the following expectation:

A2: The occurrence of identical graphemes before and after the constituent joint (e.g. in *felt+tip*) disfavors solid spelling.

In order to test Hypothesis A2, the program CompSpell counted identical letters before and after the constituents' boundaries (e.g. the two <t>'s in the middle of *felt+tip*).

Pearson's chi-square test was then conducted in SPSS for the OHS_600 compounds with the dependent variable 'compound spelling' [OHS] and the independent variable 'identical consonants or vowels across the constituent joint' [Ident_lett] (cf. Table 5.2). Since there were twenty-seven cells with expected counts below five and several counts smaller than one, the data had to be recoded in a binary comparison contrasting all compounds with repeated graphemes and those without repetition [Ident_lett_r]. The subsequent chi-square test resulted in a level of significance of $p = 0.026$ – which can be explained by the fact that the proportion of solid compounds with identical graphemes across the constituent joint is indeed lower than expected (1 hit instead of 6.3 – which corresponds to merely 5.3 per cent). This finding replicates Sepp's (2006: 115) result that identical consonants across constituent joints disfavour solid spelling. At the same time, there is an increase in open spelling (42.1 per cent) and hyphenation (52.6 per cent). Since these proportions are based on an absolute difference of two counts only, [Ident_lett_r] was considered as favouring both open and hyphenated spelling in Table A.9 in the Appendix.

Table 5.3 *Grouped identical graphemes across the constituent joint [Ident_lett_r] and spelling in OHS_600*

			Identical letter across constituent joint		Total
			–	+	
OHS	o	Count	192	8	200
		Expected Count	193.7	6.3	200.0
		% within Ident_lett_r	33.0%	42.1%	33.3%
	h	Count	190	10	200
		Expected Count	193.7	6.3	200.0
		% within Ident_lett_r	32.7%	52.6%	33.3%
	s	Count	199	1	200
		Expected Count	193.7	6.3	200.0
		% within Ident_lett_r	34.3%	5.3%	33.3%
	Total	Count	581	19	600
		Expected Count	581.0	19.0	600.0
		% within Ident_lett_r	100.0%	100.0%	100.0%

Hypothesis A2 can thus be accepted: the occurrence of identical graphemes before and after the constituent joint disfavours solid spelling. One possible explanation for this is the fact that double consonants usually signal the shortening of the preceding vowel (Swan 2005: 554): compare the short vowel in double-*<p>* *hopping* to the diphthong in single-*<p>* *hoping*. As a consequence, language users might avoid solid spellings in compounds where this could suggest a misleading shortening, e.g. of the long /ɑ:/ in uniquely hyphenated *far-reaching*. This explanation is supported by the finding that *beach-head* is the only solid compound with consonant doubling at the constituent joint: since the grapheme <ea> for the long vowel /i:/ never seems to correspond to a short /ɪ/ in English, confusion is unlikely – particularly in view of the fact that double <h> is no usual English grapheme either.

5.1.1.3 *Hypothesis A3 – Garden Path Cluster across the Constituent Joint*

The presence of garden path clusters, i.e. misleading digraphs like <ph> in *top+hat*, across the constituent joint disfavours solid spelling.
→ Refuted.

Another interesting aspect regarding graphotactics concerns the frequency of the bigrams across the constituent joint. Since bigrams across

morpheme boundaries tend to be lower in frequency than bigrams contained within morphemes, readers could use them to discover word boundaries (cf. Seidenberg 1987). This finding might explain the relative ease with which readers segment the solid compounds they encounter in a text. At the same time, it implies a certain reluctance to concatenate the constituents of compounds if lexical boundaries would be blurred, e.g. where the combination of graphemes across the constituents' boundaries would result in misleading bigraphs, such as <ph>, which usually corresponds to /f/ (e.g. in *top+heavy*). By analogy to *garden path sentences* (cf. Frazier 1987) like the classic *The horse raced past the barn fell*, such misleading clusters will be called *garden path clusters*, because they require a grapho-phonemic re-analysis. We can thus formulate the following hypothesis:

- A3: The presence of garden path clusters (i.e. misleading digraphs like <ph> in *top+bat*) across the constituent joint disfavours solid spelling.

In order to test this hypothesis, the program CompSpell automatically checked the bigrams across the constituent joints against the following list of potential garden path clusters:

<ch>	/tʃ/	<i>chin</i>
<gh>	/f/	<i>laugh</i>
<ng>	/ŋ/	<i>sing</i>
<ph>	/f/	<i>photo</i>
<sh>	/ʃ/	<i>ship</i>
<th>	/θ/ or /ð/	<i>thin, the.</i>

The items in this list correspond to Venezky's (1967: 86) simple major relational units consisting of two letters (with the exception of <ng>, which was added to the list; cf. also Venezky 1967: 90). Digraphs such as <rh> and <sc>, whose pronunciation corresponds to the value of the first letter or whose phonetic quality can be rendered by each of the two graphemes composing it, were not considered.

Altogether, OHS_600 contains fourteen garden path clusters and OHS_extra fifty-one. Pearson's chi-square test was conducted for OHS_600 with the dependent variable 'compound spelling' [OHS] and the independent variable 'misleading digraph across the constituent joint' [Digraph]. Since neither these results nor those for OHS_extra met the conditions for chi-square tests described in Moore and Notz (2006: 496),

Table 5.4 *Garden path clusters across the constituent joint [Digraph] in OHS_600 and OHS_extra*

Misleading digraph	Frequency in OHS_600	Frequency in OHS_extra
none	586	3,813
ch	1	2
gh	2	17
ng	1	8
ph	2	4
sh	0	4
th	8	16

a subsequent analysis was carried out, in which all compounds with garden path clusters were jointly contrasted with unmarked compounds [Digraph_r]. While the OHS_600 data still did not meet the requirements for statistical testing, the results for OHS_extra revealed a weak but significant effect of the variable ‘misleading digraph across the constituent joint’ ($p = 0.040$) – only reversing the expectations: the proportion of solid spellings is actually higher than usual in those compounds with misleading digraphs across the constituent joints, with twenty-four observed (as against 16.1 expected) instances and thus a proportion of 47.1 per cent as against 31.4 per cent in the unmarked compounds. As a consequence, Hypothesis A3 needs to be refuted: the presence of garden path clusters does not disfavour solid spelling but even represents a favourable influence in this respect, it seems. A possible explanation for this could be that the default reading process in solid compounds considers consonant graphemes on a one-by-one basis and only reconsiders them as graphemic combinations if the one-by-one interpretation results in an obviously incorrect analysis, e.g. in nonsensical *lights+hip* instead of actual *light+ship*. In contrast to garden path sentences, which are caused by early node closure (cf. Frazier 1987: 561–562), the unduly late node closure in garden path clusters does not seem to cause any problems that are important enough to result in the common avoidance of solid spelling.

5.1.1.4 *Hypothesis A4 – Vowel Graphemes across the Constituent Joint*

The occurrence of vowel graphemes before and after the constituent joint (e.g. in *amino+acid*) disfavours solid spelling. → Confirmed.

Table 5.5 *Grouped garden path clusters across the constituent joint [Digraph_r] and spelling in OHS_extra*

			Garden path cluster across constituent joint		Total
			–	+	
OHS_extra	o	Count	2,295	22	2,317
		Expected Count	2,286.4	30.6	2,317.0
		% within Digraph_r	60.2%	43.1%	60.0%
	h	Count	320	5	325
		Expected Count	320.7	4.3	325.0
		% within Digraph_r	8.4%	9.8%	8.4%
	s	Count	1,198	24	1,222
		Expected Count	1,205.9	16.1	1,222.0
		% within Digraph_r	31.4%	47.1%	31.6%
	Total	Count	3813	51	3,864
		Expected Count	3,813.0	51.0	3,864.0
		% within Digraph_r	100.0%	100.0%	100.0%

Many vowel grapheme combinations represent phonemes whose value cannot necessarily be derived directly from the usual pronunciation of the individual graphemes, e.g.

<e> + <a>	/e/	<i>head</i>
<e> + <i>	/ei/	<i>weight</i>
<e> + <o>	/ə/	<i>surgeon</i>

If compounds combining vowel graphemes across the constituent joint are spelled solid, a situation arises that is similar to that of the consonant garden path clusters discussed earlier, with the hypothetical spelling *tradein* (cf. also Burrige 2005: 163) incorrectly suggesting the realisation of a medial diphthong. By analogy to the discussion of garden path clusters, we may therefore formulate the following expectation:

- A4: The occurrence of vowel graphemes before and after the constituent joint (e.g. in *amino+acid*) disfavors solid spelling.

In order to test Hypothesis A4, the program CompSpell marked compounds in which any combination of the vowel graphemes <a>, <e>, <i>, <o> and <u> occurred across the constituent joint.⁴

⁴ Cf. Section 5.1.1.1 for the position-dependent consideration of <y> as a vowel grapheme.

Table 5.6 *Grouped vowel graphemes across the constituent joint [o_I_VV_2_r] and spelling in OHS_600*

			Vowel graphemes across constituent joint		Total
			–	+	
OHS	o	Count	191	9	200
		Expected Count	191.3	8.7	200.0
		% within o_I_VV_2_r	33.3%	34.6%	33.3%
	h	Count	184	16	200
		Expected Count	191.3	8.7	200.0
		% within o_I_VV_2_r	32.1%	61.5%	33.3%
	s	Count	199	1	200
		Expected Count	191.3	8.7	200.0
		% within o_I_VV_2_r	34.7%	3.8%	33.3%
Total	Count		574	26	600
	Expected Count		574.0	26.0	600.0
	% within o_I_VV_2_r		100.0%	100.0%	100.0%

OHS_600 was found to contain twenty-six compounds with vowel graphemes before and after the constituent joint [o_I_VV_2]: twenty-two sequences with double vowels (e.g. *fire+alarm*), two with three vowels (e.g. *shoo+in*) and two with four vowels (e.g. *eye+opener*). In order to meet the requirements for statistical testing, the data were recoded in a binary opposition regarding the presence/absence of vowel graphemes across the constituent joint [o_I_VV_2_r]. Pearson's chi-square test for the dependent variable 'compound spelling' and the independent variable [o_I_VV_2_r] reached a level of significance of $p = 0.001$. It becomes very clear from Table 5.6 that the compounds with vowel graphemes across the constituent joint avoid solid spelling (with the sole exception of *firearm* resulting in a minute proportion of 3.8 per cent) and that compounds in this category are mostly hyphenated (61.5 per cent as against 32.1 per cent in the unmarked sample). Hypothesis A4 can therefore be confirmed: the occurrence of vowel graphemes before and after the constituent joint disfavours solid spelling. This result confirms Sepp's (2006: 115) observation, which is based on a wide vowel concept encompassing both spelling and pronunciation (and thus recognising final vowels both in *snow* and in *middle*), that vowel-vowel sequences are avoided at the internal boundary in solid forms. Since no similar effect was found for consonant graphemes (cf. Hypothesis A3), the findings suggest that the reading of vowel

graphemes proceeds in a different way than that of consonant graphemes. A possible explanation for this is that combinations of vowel graphemes almost exclusively result in a change of the vowel quality (compared to the pronunciation of the individual vowel graphemes), while this is very rarely the case for consonants.

5.1.2 Capitalisation

Context-independent (i.e. non-sentence-initial) capitalisation is discussed in the literature as exerting some influence on English compound spelling.

5.1.2.1 Hypothesis A5 – Capitalisation of Second Constituent

The capitalisation of a compound's second constituent (e.g. in *reality+TV*) disfavours solid spelling. → Tentatively confirmed.

There seems to be a certain reluctance to concatenate combinations containing proper names (which are capitalised by definition). Thus Quirk et al. (1985: 1569) observe a tendency in proper names which are preceded by a combining form (e.g. *Sino-Russian* or *Anglo-American*) away from solid spelling, because it is inhibited by the initial capital of the second constituent (Quirk et al. 1985: 1569). Fowler (1921: 12) notes:

When a hyphenated word becomes sufficiently current and familiar to cast its hyphen, it must also cast the capital of its second component, if it had one in its hyphenated form, e.g. *Home-Rule* would become *Homerule*. The impropriety of writing a capital letter in the middle of a word will thus sometimes forbid coalescence.

As the partial capitalisation in compounds containing acronyms like *IP address* (cf. the *Macmillan English Dictionary for Advanced Learners* 2007) also seems to inhibit solid spelling, we may formulate the following combined expectation:

A5: The capitalisation of a compound's second constituent (e.g. in *reality+TV*) disfavours solid spelling.

In order to test Hypothesis A5, capitalisation [Cap] was coded manually for all Master List compounds by carrying out case-sensitive searches of all letters of the alphabet.

The OHS_600 list contains no capitalised compounds, and the larger OHS_extra list only contains six such items (*all-American*,

doubting Thomas, *girl Friday*, *middle C*, *pay TV*, *reality TV*). While none of these is spelled solid – which supports the hypothesis – and while all except the hyphenated first compound are spelled open, the number is far too small to permit any statistically valid conclusions. Master_5+ yielded the biconstituent compound *double+Dutch*, which varies between open spelling (four times) and hyphenation (once), and the triconstituent compound *plaster of Paris* (always spelled open at both constituent joints). While Hypothesis A₅ can neither be accepted nor refuted for lack of statistical backing, there is thus an obvious tendency towards open spelling, a slightly less strong tendency towards hyphenation and a tendency to avoid solid spelling in compounds with capitalised non-initial constituents. Note, however, that the orthographic feature of capitalisation is inseparable from part of speech (since names are generally nouns), semantics (since proper nouns are characterised by having reference but no meaning; cf. Lyons 1977: 219) and word formation type (when the capitalised constituents are acronyms), so that a particular spelling behaviour may actually be attributable to these interrelations.

5.1.3 Symbols, Numbers and Punctuation Marks

Prototypical compounds consist only of letters and possibly spaces or hyphens. As a consequence, the presence of other punctuation marks, symbols, numbers or typographic modifications in more unusual compounds alters their visual appearance and may have repercussions on their spelling. For instance, the *GPO Style Manual* (2008: 79) advocates that one should “not print a hyphen in a unit modifier containing a letter or a numeral as its second element”, e.g. *grade A milk* and *point 4 program*. Chemical elements containing numbers are said to be preferably hyphenated (e.g. *Freon-12*), except if superior figures are used (*Sr⁹⁰*; cf. *GPO Style Manual* 2008: 82).

Punctuation marks within compounds are unusual but do exist: thus two compounds from the original LDOCE compound list contained a comma (*all-singing*, *all-dancing* and *two up*, *two down*). As the proper name of the musical *Oklahoma!* comprises an exclamation mark (Huddleston and Pullum 2002: 1759), it is possible to find occurrences of *Oklahoma! cast* and *Oklahoma! plot* with a compound-internal exclamation mark (Google July 2017). The most common compound-internal punctuation mark beside the hyphen is the apostrophe.

5.1.3.1 Hypothesis A6 – Compound-Internal Apostrophe

The occurrence of an apostrophe within a compound (e.g. in *seller's+market*) disfavours solid spelling. → Confirmed.

The spelling of so-called *genitive compounds*, i.e. noun+noun compounds whose first constituent contains a genitive, has been observed to depend on the presence or absence of an apostrophe: according to Merriam-Webster (2001: 101–102), genitive compounds containing an apostrophe are usually spelled open (*seller's market*) and sometimes hyphenated (*bull's-eye*), whereas genitive compounds which have lost their apostrophe are usually spelled solid (*menswear*). This may be tentatively explained by the fact that apostrophes used in genitives are conventionally either followed by the genitive {S} and then by a space (*Joan's house*) or directly by a space (*the Joneses' car*). The solid combination of constituent + apostrophe + genitive {S} + constituent (**seller'smarket*) would violate this convention and blur the distinction from a less frequent use of the apostrophe to mark omitted letters (e.g. <o> in *don't* or <v> in the Master List compound *ne'er-do-well*). All of this seems to indicate a relatively strong tendency not to concatenate compounds if they contain some element that is not a letter. Since apostrophes were the only sufficiently frequent punctuation mark in the dataset, the hypothesis was formulated in the following way:

A6: The occurrence of an apostrophe within a compound (e.g. in *seller's+market*) disfavours solid spelling.

Apostrophes were coded with the code *a* for all Master List compounds in a separate column. A further differentiation was introduced by considering position with the codes

- *a* for apostrophes on the first constituent (e.g. *potter's+wheel*)
- *a2* for apostrophes on the second constituent (e.g. *blind+man's+buff*)
- *a2a* for apostrophes around the second constituent (e.g. *surf+n'+turf*).

Only two OHS_600 compounds contain apostrophes [Apostr], and both are spelled open in all the dictionaries (*banker's+order* and *cat's+cradle*). OHS_extra contains sixteen compounds with apostrophes; all of them noun + genitive {S} (sometimes combined with plural) + noun, and all of them using open spelling too (*athlete's foot*, *baker's dozen*, *collector's item*, *crow's feet*, *devil's advocate*, *director's cut*, *farmers' market*, *fool's gold*, *guard's van*, *ladies' man*, *men's room*, *saint's day*, *shepherd's pie*, *traveller's cheque*, *women's studies*, *writer's cramp*). In

Table 5.7 *Apostrophes [Apostr] and spelling in OHS_extra*

			Apostrophe		Total
			-	+	
OHS_extra	o	Count	2,301	16	2,317
		Expected Count	2,307.4	9.6	2,317.0
		% within Apostr	59.8%	100.0%	60.0%
	h	Count	325	0	325
		Expected Count	323.7	1.3	325.0
		% within Apostr	8.4%	0.0%	8.4%
	s	Count	1,222	0	1,222
		Expected Count	1,216.9	5.1	1,222.0
		% within Apostr	31.8%	0.0%	31.6%
Total	Count		3,848	16	3,864
	Expected Count		3,848.0	16.0	3,864.0
	% within Apostr		100.0%	100.0%	100.0%

spite of the small number, this sample meets the requirements for statistical testing. Pearson’s chi-square test for the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘presence of an apostrophe’ [Apostr] results in a high level of significance ($p = 0.005$). We may therefore conclude that Hypothesis A6 is supported by the data and that the occurrence of an apostrophe within a biconstituent compound disfavours solid spelling.

Note, however, that the Master_5+ compounds with more than two constituents which contain no genitive behave slightly differently (cf. Table 5.8): *jack+o’+lantern* and *will+o’+the+wisp* (in which the apostrophe signals the omission of preposition-final <f>) also have hyphenated variants. *Rock+’n’+roll* and *drum+’n’+bass*, which contain two apostrophes each, even have completely solid variants – possibly due to the special situation that the two apostrophes reduce the middle constituent from both sides. These results suggest that the usual difference in spelling behaviour between genitive apostrophes (which are followed by a space immediately after the apostrophe for the plural or after the {S} morpheme for the singular) and omission apostrophes (which are surrounded by letters, as in *cap’n*) also extends to compound spelling and influences the selection of spelling variants. For biconstituent compounds, we may, however, conclude that the presence of an apostrophe makes open spelling extremely likely and solid spelling extremely unlikely.

Table 5.8 *Master_5+ compounds with more than two constituents which contain apostrophes but no genitive*

	OO	OOO	HH	HHH	HO	HOH	SS
<i>drum+ 'n'+bass</i>	3	0	0	0	0	0	2
<i>jack+o'+lantern</i>	0	0	4	0	1	0	0
<i>rock+ 'n'+roll</i>	4	0	0	0	0	0	1
<i>will+o'+the+wisp</i>	0	2	0	2	0	1	0

5.1.4 Summary

Section 5.1 investigates spelling-related variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Number of consonant graphemes across constituent boundaries
- Identical graphemes before and after constituent boundaries (*glow+worm*)
- Garden path clusters: misleading digraphs across constituent boundaries (in *ant+hill*)
- Vowel graphemes before and after constituent boundaries (*amino+acid*)
- Capitalisation of one or more constituents (*three+R's*)
- Occurrence of one or more apostrophes within the compound (*seller's+market*).

Some of the findings contradict the expectations:

- Consonant clusters with four or more graphemes across the constituent joint (*grind+stone*) do not disfavour solid spelling. [A1] Since the database contained only few compounds with five or more consonant graphemes in a row, the results are possibly only representative for clusters of four consonants across the constituent joint.
- Garden path clusters across the constituent joint (e.g. <th> in *ant+hill*) do not disfavour solid spelling but actually favour it. [A3]

We do, however, find the expected effect of several other variables:

- Identical graphemes across the constituent joint disfavour solid spelling. [A2]
- Vowel graphemes across the constituent joint disfavour solid spelling. [A4]

- The presence of an apostrophe makes open spelling extremely likely and solid spelling extremely unlikely. [A6]

In addition, the analyses revealed the following statistically significant observation:

- Compounds with vowel graphemes across the constituent joint favour hyphenation. [A4]

Yet another finding lacks statistical backing but is indicative of a clear tendency:

- Compounds with capitalised non-initial constituents have a strong tendency towards open spelling, a slightly less strong tendency towards hyphenation and a tendency to avoid solid spelling. [A5]

There is also a counter-intuitive additional finding:

- Consonant clusters with four or more graphemes across the constituent joint disfavour open spelling. [A1]

5.2 Length

With regard to length as a possible determinant of compound spelling, Bauer (1998: 69) claims that long words tend towards open spelling, while short words are more often spelled solid. Since length can be measured in different ways, namely by number of constituents, syllables or letters, the following subsections discuss measurement-specific length-related hypotheses for whole compounds or their constituents (cf. 5.2.9 for a summary).

5.2.1 Hypothesis B1 – Three or More Constituents

Compounds containing three or more constituents disfavour solid spelling. → Tentatively confirmed.

The largest measure for a compound's length is the number of constituents. It may differ from the number of morphemes in that a compound's base may be a prefixation or suffixation or even a compound. The program CompSpell determined the number of constituents [Constituents] by counting the number of <+> signs that had previously been inserted at each compound's constituent joints and by adding one (e.g. *ball+room+dancing* contains two <+> signs and three constituents). Strings containing more than "about five elements" are very unusual in English (Bauer 2009:

350), because “limitations on short-term memory may affect the length of compounds in actual use” (Bauer 1983: 67). Marchand (1960b: 412) automatically interprets combinations of four constituents as syntactic groups, and Schmid (2011: 205–206) argues that extremely complex compounds – such as *holiday car sightseeing trip* – are relatively rare in spoken English, because the realisation with unit intonation and main stress requires too much planning ahead and places too high a demand on processing. As a consequence, one would rather expect ‘*and then we went on a sightseeing trip with our holiday car*’ in spoken language. Still, compounds with even more constituents can be found; often ad hoc compounds fulfilling an adjectival function with a tendency towards hyphenation, such as *I-don’t-care-what-you-do (attitude)*. Merriam-Webster (2001: 104) explicitly recognises noun compounds with four constituents and states that these can be hyphenated or open, and Inhoff, Radach and Heller (2000: 24) note that “English rarely – if ever – contains concatenated compounds with more than two constituent words”. To sum up, the literature discusses a general tendency for compounds with many constituents that can be formulated as Hypothesis B1:

B1: Compounds containing three or more constituents disfavour solid spelling.

Hypothesis B1 was tested in Master_5+, since this list also contains compounds with more than two constituents (282 types with three constituents and 23 types with four constituents, e.g. *jack+in+the+box* or *up+to+the+minute*). The total number of hits in all dictionaries for these long compounds is 1,638. If solid spelling is defined in the strictest possible way as exclusive category membership in the categories SS (= two solid constituent joints) or SSS, this requirement is only met by the triconstituent compound *high+way+man*. If the definition of solid spelling is extended to comprise compounds with one or more dictionary hits in the categories SS or SSS, only four compounds with three constituents meet the condition (and not a single one with four). Of these four compounds, *drum’n’bass* and *rock’n’roll* contain two apostrophes each, which signal the omission of letters and segment the letter sequence into three parts, so that no additional spaces or hyphens are required. The other two solid compounds with three constituents are *highwayman* and *nightwatchman*, both of which are special in that their final constituent *man* has a rather suffix-like character (cf. 5.5.1). Among the compounds with a frequency between three and four in the Master List, *aircraftman*, *newspaperman* and *railwayman* are also spelled solid. Since these compounds are composed of three

consecutive nouns, this finding would seem to contradict Sepp's (2006: 26) result that "English limits concatenation of nouns to just two". However, the small number of instances in the present dataset, combined with the special status of the compound-final constituent, conveys exception status to this group rather than really challenging Sepp's results. If the solid category is defined even less strictly as comprising combinations of S with other spellings, the 1,196-compound list Master_5+ contains two HS types (e.g. *daddy-long+legs*), twenty-seven OS types (e.g. *fairy god+mother*), thirty SO types (e.g. *gold+fish bowl*), one OOS type (*local area net+work*), one SOO type (*day+light saving time*) and one SHO type (*day+light-saving time*). The consideration of the examples in parentheses will suffice, however, to convey the idea that such a definition would stretch the concept of solid spelling too much, and we can conclude that solid spelling is found relatively rarely in long compounds. Hypothesis B1 is thus confirmed by the data: compounds containing three or more constituents disfavour solid spelling.

5.2.2 Hypothesis B2 – Four or More Syllables

Compounds containing four or more syllables disfavour solid spelling. → Confirmed.

The second largest unit in which the length of compounds can be measured is the syllable. CompSpell's automated syllable count derives this phonological property from the compounds' written form by using the number of vowels as an approximation for the number of syllables and by following various syllabification principles partly inspired by Hoover (1971) and Venezky (1970), which were extended and refined in several test cycles (cf. Table 5.9).

A manual check of 100 random compounds found that the syllabification principles in Table 5.9 yielded correct results for 94 per cent of the test items. Since the syllabification of some constituents appears to be idiosyncratic and cannot be derived systematically from their written form (e.g. <ea> in *reach* vs. *react*; cf. Hoover 1971: 159), the principles were deemed sufficiently precise for the purpose of the present study.

Following the syllabification principles outlined earlier, CompSpell determined that the number of syllables of the OHS_600 compounds lies between two and nine, with the median at two.

Table 5.9 *CompSpell's syllabification principles*

-
-
- The single vowel graphemes <a>, <e>, <i>, <o> and <u> correspond to one syllable core each and thus count as one syllable each.
 - Two consecutive vowel graphemes correspond either to a long vowel (*root*) or to a diphthong (*loud*) and thus to a single syllable.
 - The sequence <eau> (*beauty*), which consists of three vowel graphemes, counts as one vowel and thus one syllable.
 - The letter <y> can represent either a vowel or a consonant. Before a consonant (*byte*, *boys*) or a constituent boundary (*my*, *prey*), it counts as a vowel, regardless of the preceding grapheme.
 - Before a vowel grapheme (*you*), <y> is considered a consonant, except if the <y> is preceded by yet another vowel (*eye*, *soya*). In this context <y> counts as a vowel. However, when <y> is preceded by a vowel and followed by <i> (*crop+spraying*), the <i> usually belongs to the suffix *-ing* (which is realised as a separate syllable), so that <y> counts as a consonant in this context.
 - At the end of a constituent (i.e. before a space, hyphen, + or the cell frame), <e> is usually silent (*snake+bite*). It therefore only counts as a vowel if the constituent contains no other vowel (*the*) or if the <e> is preceded by <l> (*subtle*).
 - At the end of a word, <ed> is usually a suffix and pronounced /d/ or /t/ depending on the voicing of the context (*loved*, *packed*), i.e. in a non-syllabic manner (Swan 2005: 393–394). Constituent-final <e> followed by <d> therefore only counts as a syllabic vowel if preceded by <t> (*sharp+witted*) or <d> (*simple+minded*) and in constituents that contain no other vowel (*bed*).
 - At the end of a word, <es> is often a plural marker (*sales*) pronounced in a non-syllabic manner (Swan 2005: 518–519). It only counts as syllabic vowel if preceded by <s>, <z>, <sh>, <ch> or <dg> (e.g. *glasses*, *breeches*) and in constituents that contain no other vowel (*yes*).
 - At the end of a constituent, <ier> counts as two syllables, because the <er> usually corresponds to the comparative suffix *-er*, which is realised as a separate syllable (*holier+than+thou*).
 - If a constituent contains no vowel grapheme (e.g. *x* or *ff*), it is assigned a default value of one syllable, because a vowel is inserted when reading out compounds such as *x+axis* or *ff+word* (possibly as a phonological reflection of their status, since compounds necessarily comprise at least two constituents and thus two syllables).
-
-

In order to test whether long compounds avoid solid spelling, it was planned to operationalise short, intermediate and long compounds (measured in syllables) as the top 25 per cent, middle 50 per cent and bottom 25 per cent of a length-ordered list of compound types in OHS_600, but the distribution (cf. Table 5.10) made it necessary to draw the line at four syllables for the long compounds, to avoid making two groups of 50 per cent each. Hypothesis B2 can thus be formulated more precisely as:

Table 5.10 *Number of syllables of the whole compound*
[Syll_total] in OHS_600

Number of syllables	Frequency	Per cent
2	316	52.7
3	193	32.2
4	65	10.8
5	20	3.3
6	4	.7
7	1	.2
9	1	.2
Total	600	100.0

B2: Compounds containing four or more syllables disfavour solid spelling.

To meet the requirements for statistical testing, the OHS_600 compounds with four or more syllables had to be treated jointly [Syll_total_r]. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'recoded number of syllables' [Syll_total_r] reached a value of $p = 0.000$. The number of syllables is thus highly significant for the spelling of biconstituent English compounds. It is immediately obvious from Table 5.11 that this is indeed due to the complete avoidance of solid spelling in the compounds comprising more than three syllables, which is counterbalanced by a drastic preference for open spelling (73.6 per cent) in that group. This result is in line with Rakić's (2009: 62) observation that "compounds with more than three syllables are overwhelmingly written open" (cf. also Rakić 2010) and Sepp's (2006: 88) assumption that "there is a limit on the number of syllables allowed in a closed compound".⁵ Hypothesis B2 can thus be confirmed: compounds containing four or more syllables disfavour solid spelling. In addition, we can observe that the compounds in the shortest group behave in direct opposition, with a strongly reduced amount of open spelling (14.2 per cent) and a predominance of solid spelling (52.8 per cent). The compounds with three syllables, by contrast, already behave like those with four or more syllables, with an increase in

⁵ Note, however, that both Rakić and the present study focus on biconstituent compounds. Since phrase compounds as particularly long compounds are predominantly hyphenated, one may expect some turning point for this group of compounds.

Table 5.11 *Grouped number of syllables of the whole compound [Syll_total_r] and spelling in OHS_600*

			Number of syllables of compound			Total
			2 (= short)	3	4-9 (= long)	
OHS	o	Count	45	88	67	200
		Expected Count	105.3	64.3	30.3	200.0
		% within Syll_total_r	14.2%	45.6%	73.6%	33.3%
	h	Count	104	72	24	200
		Expected Count	105.3	64.3	30.3	200.0
		% within Syll_total_r	32.9%	37.3%	26.4%	33.3%
	s	Count	167	33	0	200
		Expected Count	105.3	64.3	30.3	200.0
		% within Syll_total_r	52.8%	17.1%	0.0%	33.3%
	Total	Count	316	193	91	600
		Expected Count	316.0	193.0	91.0	600.0
		% within Syll_total_r	100.0%	100.0%	100.0%	100.0%

the number of open spellings and a decrease in the number of solid spellings – only with a less marked difference. This very interesting finding may explain some of the difficulties language users have with English compound spelling, because what superficially appears to be a continuum is actually none. Instead, there is a clear grouping into bisyllabic compounds and others, which is statistically backed by a chi-square test ($p = 0.000$) carried out on recoded data, for the dependent variable ‘compound spelling’ [OHS] and the independent variable [Syll_total_rr], for which the OHS_600 compounds were grouped into bisyllabic compounds vs. compounds with three or more syllables. Since compounds with three syllables are superficially similar to compounds with two syllables, language users may not expect such a completely different spelling behaviour.

5.2.3 Hypothesis B3 – Eleven or More Letters

Compounds containing eleven or more letters disfavour solid spelling. → Confirmed.

The shortest unit in which the length of compounds can be measured is the number of letters. The program CompSpell automatically counted the number of letters of the whole compound [Lett_total]. The length of

Table 5.12 *Number of letters of the whole compound [Lett_total] in OHS_600*

No of letters	Frequency	%
4	1	0.2
5	8	1.3
6	26	4.3
7	77	12.8
8	97	16.2
9	102	17.0
10	100	16.7
11	77	12.8
12	51	8.5
13	26	4.3
14	14	2.3
15	10	1.7
16	5	0.8
17	1	0.2
18	3	0.5
20	1	0.2
22	1	0.2
Total	600	100.0

the OHS_600 compounds varies between four and twenty-two letters, with the median at nine.

Again, the cut-off point for the group of the longest compounds was determined by the length of the shortest words still included in the 25 per cent of the longest compound types in OHS_600, which corresponds to compounds with eleven or more letters. The hypothesis was then formulated more specifically as:

B3: Compounds containing eleven or more letters disfavour solid spelling.

In order to meet the requirements for statistical testing, the data were recoded by considering all long compounds as a single value [Lett_total_r] and by treating all compounds with four to eight letters jointly. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped number of letters of the whole compound' [Lett_total_r] for OHS_600 suggests that the number of letters is highly significant for the spelling of biconstituent English compounds ($p = 0.000$). Only 4.8 per cent solid spellings in the

Table 5.13 *Grouped number of letters of the whole compound [Lett_total_r] and spelling in OHS_600*

			Number of letters of compound				Total
			4–8	9	10	11–22	
OHS	o	Count	33	21	32	114	200
		Expected Count	69.7	34.0	33.3	63.0	200.0
		% within Lett_total_r	15.8%	20.6%	32.0%	60.3%	33.3%
	h	Count	80	26	28	66	200
		Expected Count	69.7	34.0	33.3	63.0	200.0
		% within Lett_total_r	38.3%	25.5%	28.0%	34.9%	33.3%
	s	Count	96	55	40	9	200
		Expected Count	69.7	34.0	33.3	63.0	200.0
		% within Lett_total_r	45.9%	53.9%	40.0%	4.8%	33.3%
	Total	Count	209	102	100	189	600
		Expected Count	209.0	102.0	100.0	189.0	600.0
		% within Lett_total_r	100.0%	100.0%	100.0%	100.0%	100.0%

longest compounds (nine counts instead of the expected sixty-three), compared to 60.3 per cent open spellings in this group, speak a clear language. Hypothesis B3 is thus supported by the data: long compounds measured in number of letters disfavour solid spelling. This is accompanied by a trend towards solid and away from open spelling in the very short compounds. It is interesting to note that solid spelling (which represents the shortest orthographic variant) is used with the shortest constituents, whereas open and hyphenated spellings (which are longer in terms of the space they use) link longer constituents. All this would seem to suggest that solid spelling confers visual weight to shorter compounds, thereby making them more easily identifiable as units, whereas non-solid spelling is an advantage for longer biconstituent compounds, because it facilitates the segmentation process (cf. 1.1.1).

5.2.4 Hypothesis B4 – Single-Letter Constituent

Compounds containing one constituent consisting of a single letter (e.g. *T+shirt*) disfavour solid spelling. → Tentatively confirmed.

Most commonly, grammars and style guides recommend the hyphenation of compounds with an initial single-letter constituent (e.g. *T-shirt*), but open spelling is also recorded, e.g. in *Tsquare*, *B vitamin* (Quirk et al. 1985: 1613–1614; Merriam-Webster 2001: 103; Butcher 1992: 13). Compounds

whose second constituent consists of a single letter, such as the (hyphenated) adjective *Jackie-O*, ‘fashionable in the style of Jacqueline Onassis’ (Macmillan 2007 s.v. *Jackie-O*), are very rare. The single-letter constituents in compounds with solid spelling commonly seem to have developed an affix-like status (e.g. *e+mail*). Taking everything into account, we may therefore formulate the following expectation:

B4: Compounds containing one constituent consisting of a single letter (e.g. *T-shirt*, *x ray*) disfavour solid spelling.

Since OHS_600 contains no compounds with single-letter constituents, OHS_extra was analysed – but this larger sample also only yielded *middle C* (with a compound-final single letter and exclusively open spelling). The consideration of Master_5+ yielded no one-letter constituent in the initial position either – only compound-medially in *chock-a-block*, *cock-a-hoop* and *cock-a-leekie* (all with a single letter as the second constituent and exclusively hyphenated spelling at both joints) and *son of a bitch* (with the single letter in the third place and completely open spelling in all five hits). A tendency to avoid solid spelling can thus be observed regardless of the position of the briefest possible constituent – even if the compounds with *a* differ from the others in that here the single-letter constituent is a determiner and no abbreviation, as usual. A possible explanation for the general reluctance to concatenate may be that the hyphen (or space) in the more prototypical single-letter constituent compounds also signals a change in pronouncing conventions: while <T> is usually pronounced /t/ within English lexemes, it is realised as /ti:/ in *T-shirt*; i.e. as the letter of the alphabet singled off in the spelling. Yet while we can observe a tendency for Hypothesis B4 to be confirmed by the data, it lacks statistical backing, since the database is not sufficiently large to permit statistical testing.

5.2.5 Hypothesis B5 – Length Difference between Constituents (Letters)

A ratio exceeding 2:1:1:2 in the number of letters of the constituents of biconstituent compounds disfavours solid spelling. → Confirmed.

One expectation in the context of the present study was that spellers are reluctant to concatenate very dissimilar constituents (cf. Hypothesis L2), e.g. with regard to length. Since constituent length can be measured using either the number of syllables or the number of letters, the following more specific hypothesis can be formulated:

B5: A large difference in the number of letters of the constituents disfavours solid spelling.

A large difference was defined as a ratio exceeding 2:1 or 1:2 in compounds with two constituents. The directionality of the differences (i.e. whether the first or second constituent is the longer one) was not considered. Since the ratio of biconstituent compounds such as *AB* was calculated by performing the division *A/B* in *Excel*, the values obtained were either

1. smaller than one if A is shorter than B
2. larger than one if A is longer than B
3. equal one if A is identical with B.

The cut-off point for compounds with a large difference between the constituents was consequently 0.5 for the first group and 2.0 for the second group. The cut-off points themselves were not included. As a consequence, *dragon+fly* with a ratio of 2.0 was not considered a compound with a large length difference between the constituents, whereas *suspender+belt* with a ratio of 2.25 was a successful candidate.

The analysis revealed that 55 of the 600 OHS_600 compounds combine constituents of very different length and display a ratio exceeding 2:1 or 1:2 regarding the number of letters [Lett_diff_12], e.g. *personal+ad* with a ratio of 4:1. Pearson's chi-square test for the dependent variable 'compound

Table 5.14 *Compounds whose constituents exceed a ratio of 2:1/1:2 (in number of letters) [Lett_diff_12] and spelling in OHS_600*

			Length difference (letters)		Total
			–	+	
OHS	o	Count	188	12	200
		Expected Count	181.7	18.3	200.0
		% within Lett_diff_12	34.5%	21.8%	33.3%
	h	Count	158	42	200
		Expected Count	181.7	18.3	200.0
		% within Lett_diff_12	29.0%	76.4%	33.3%
	s	Count	199	1	200
		Expected Count	181.7	18.3	200.0
		% within Lett_diff_12	36.5%	1.8%	33.3%
Total	Count		545	55	600
	Expected Count		545.0	55.0	600.0
	% within Lett_diff_12		100.0%	100.0%	100.0%

spelling' [OHS] and the independent variable 'ratio of the number of letters of the constituents exceeds 2:1/1:2' [Lett_diff_12] returns a highly significant ($p = 0.000$) result for the OHS_600 compounds. While the 545 unmarked cases show an almost equal distribution of open, hyphenated and solid spellings with a proportion of roughly one-third each, the marked compounds behave very differently, with the almost complete avoidance of solid spelling (the single exception is *weather+man*, whose second constituent is a quasi-suffix) and a clear predominance of hyphenation (42 instead of 18.3 expected cases = 76.4 per cent). Hypothesis B₅ can therefore be confirmed: a large difference in the number of letters of the constituents results in the avoidance of solid spelling. This seems to suggest that in compound reading, default segmentation is expected roughly in the middle of compounds, so that spellers tend to support the readers in the case of compounds which do not conform to this pattern by spelling such compounds with a clear indication at the constituent joint (cf. also 5.2.9) – a hypothesis which could be tested in future research.

5.2.6 Hypothesis B₆ – Length Difference between Constituents (Syllables)

A ratio exceeding 2:1/1:2 in the number of syllables of the constituents of biconstituent compounds disfavours solid spelling. → Confirmed.

By analogy to Hypothesis B₅, it was investigated whether compound spelling is influenced by differences in the constituents' number of syllables:

B₆: A large difference in the number of syllables of the constituents disfavours solid spelling.

Hypothesis B₆ was tested by analogy to Hypothesis B₅, using the syllabic counts described in 5.2.2. Pearson's chi-square test was carried out for OHS_600 with the dependent variable 'compound spelling' [OHS] and the independent variable 'ratio of the number of syllables of the constituents exceeds 2:1/1:2' [Syll_diff_12]. The results are highly significant ($p = 0.000$): while the distribution of the unmarked cases corresponds to the expected proportion of about one-third for each spelling variant, the thirty-one compounds with an important length difference between the constituents (e.g. *security+risk* with a ratio of 4:1) behave very differently: they are never spelled solid, but open spelling is clearly increased (21 instead of the 10.3 expected cases = 67.7 per cent). Hypothesis B₆ can therefore be accepted: a ratio exceeding 2:1/1:2 in the number of syllables of

Table 5.15 *Compounds whose constituents exceed a ratio of 2:1:1:2 (in number of syllables) [Syll_diff_12] and spelling in OHS_600*

			Length difference (syllables)		Total
			–	+	
OHS	o	Count	179	21	200
		Expected Count	189.7	10.3	200.0
		% within Syll_diff_12	31.5%	67.7%	33.3%
	h	Count	190	10	200
		Expected Count	189.7	10.3	200.0
		% within Syll_diff_12	33.4%	32.3%	33.3%
	s	Count	200	0	200
		Expected Count	189.7	10.3	200.0
		% within Syll_diff_12	35.1%	0.0%	33.3%
Total	Count		569	31	600
	Expected Count		569.0	31.0	600.0
	% within Syll_diff_12		100.0%	100.0%	100.0%

the constituents of biconstituent compounds disfavours solid spelling. The explanation is the same as for Hypothesis B₅, namely that default segmentation seems to be expected roughly in the middle of compounds and that spellers tend to support the reading of more unusual compounds by using clear segmentation cues.

5.2.7 Hypothesis B₇ – Three-Letter Constituent(s)

Compounds containing one or more constituents comprising three letters (e.g. *pig+sty*) favour solid spelling. → Refuted for the first constituent and confirmed for the second constituent.

Word length and part of speech correlate to a certain extent, with the shortest English words tending to be high-frequency grammatical words (e.g. *I*, *it*, *be* or *do*) and long words tending to be lexical words (e.g. adjectives such as *beautiful*). Compound constituents consisting of a single letter are often abbreviations (cf. 5.2.4), whereas constituents with two letters seem to be mostly pronouns or prepositions (e.g. *me*, *on*). Sporadic observations in the compilation of the Master List suggested that constituents with three letters may also behave in a special way and influence compound spelling: thus LDOCE spells *gear stick* and *gear shift* (with four-letter constituents each) with a space, whereas *gearbox* with a three-letter second constituent is spelled solid. This may be indicative of a tendency to

use solid spelling in compounds with (at least) one constituent consisting of three letters (e.g. *pig+sty* or *mud+slide*) and is investigated in Hypothesis B7:

B7: Compounds containing one or more constituents comprising three letters (e.g. *pig+sty*, *mud+slide*) favour solid spelling.

Separate tests were carried out for the first and the second constituent. Pearson's chi-square test yielded a significant correlation between the independent variable 'number of letters of the first constituent' [Lett_1_r] (with all instances of nine or more letters treated jointly in order to meet the requirements for statistical testing) and 'compound spelling' [OHS] for the OHS_600 compounds ($p = 0.000$). Nonetheless, Hypothesis B7 has to be rejected for the first constituent: with a proportion of 33.3 per cent, compounds containing a first constituent comprising three letters do not favour solid spelling. The statistical significance results from another correlation: the longer the first constituent, the more likely the compound is to use open spelling.⁶ The compounds with the shortest first constituents (including those with three letters) are mainly hyphenated and rarely spelled open.⁷ The turning point in the development comes with a length of five letters, when all spelling variants are about equally probable.

By analogy, Pearson's chi-square test was carried out on the OHS_600 compounds for the dependent variable 'compound spelling' [OHS] and for the independent variable 'number of letters of the second constituent' [Lett_2_r] (with all instances of nine or more letters treated jointly again in order to meet the requirements for statistical testing). The result was highly significant ($p = 0.000$), but the length of the second constituent has a very different effect on the spelling compared to the length of the first constituent: while the shortest second constituents are overwhelmingly hyphenated (93.3 per cent), second constituents with a length of three to five letters clearly make a compound tend towards solid spelling, as expected by Hypothesis B7. As soon as a plateau of six letters is reached, solid spelling becomes practically non-existent and open and hyphenated spellings are usually about equally likely. Taking the data for both constituents into

⁶ The slight numerical decrease in the last column is due to a statistical outlier among the five ten-letter constituents. Otherwise, all compounds with a first constituent measuring between nine and fourteen letters are always spelled open and never hyphenated.

⁷ As expected, the proportion of four verbs and eleven grammatical parts of speech of the first constituent is particularly high among the sixteen items with a two-letter first constituent (e.g. *go+ahead*, *be+man*, *no+frills*).

Table 5.17 Grouped number of letters of second constituent [Lett_2_r] and spelling in OHS_600

[illegible]

account, we may conclude that Hypothesis B7 has to be refuted for the first constituent but can be accepted for the second one: compounds whose second constituent comprises three letters favour solid spelling. Furthermore, Kuperman and Bertram's (2013: 953) finding that "[l]onger constituents came with a preference for spacing, rather than hyphenation" can only be confirmed for the first constituent but not for the second one.

5.2.8 Hypothesis B8 – Monosyllabic Constituent(s)

Compounds containing one or more monosyllabic constituents (e.g. *shoe+string*) favour solid spelling. → Tentatively confirmed.

While Hypothesis B7 attributed the solid spelling of *gearbox*, *pigsty* etc. to the presence of one or more three-letter constituents, an alternative explanation (which could also explain the preference for the solid spelling of *shoestring*) would be that these compounds contain at least one monosyllabic constituent. This can be formulated as Hypothesis B8:

B8: Compounds containing one or more constituents comprising one syllable (e.g. *shoe+string*) favour solid spelling.

Again, separate tests were carried out for the first and the second constituents. In order to meet the requirements for statistical testing, the number of syllables of the first constituent of the OHS_600 compounds was recoded by treating instances with three and four syllables jointly [Syll_1_r]. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped length of the first constituent' [Syll_1_r] yielded a highly significant result ($p = 0.000$).

While the monosyllabicity of the first constituent does indeed favour solid spelling (173 instead of 141.3 expected counts = 40.8 per cent), it favours hyphenation almost equally much (38.7 per cent). As a consequence, Hypothesis B8 can be accepted for the first constituent – but it would have to be refuted in a strict reading which understands favouring as a clear preference over both alternative spelling variants. Note also that bisyllabic first constituents still result in a certain amount of hyphenation (23.1 per cent), but that there is a clear tendency for first constituents of more than one syllable to correlate with open spelling (58.5 per cent).

The recoded number of syllables of the second constituent [Syll_2_r] treats all values larger than two jointly. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent

Table 5.18 *Grouped length of first constituent (in syllables) [Syll_1_r] and spelling in OHS_600*

			Length of first constituent (syllables)			Total
			1	2	3-4	
OHS	o	Count	87	86	27	200
		Expected Count	141.3	49.0	9.7	200.0
		% within Syll_1_r	20.5%	58.5%	93.1%	33.3%
	h	Count	164	34	2	200
		Expected Count	141.3	49.0	9.7	200.0
		% within Syll_1_r	38.7%	23.1%	6.9%	33.3%
	s	Count	173	27	0	200
		Expected Count	141.3	49.0	9.7	200.0
		% within Syll_1_r	40.8%	18.4%	0.0%	33.3%
Total	Count		424	147	29	600
	Expected Count		424.0	147.0	29.0	600.0
	% within Syll_1_r		100.0%	100.0%	100.0%	100.0%

variable [Syll_2_r] is highly significant ($p = 0.000$). Monosyllabic second constituents clearly favour solid spelling (44.9 per cent) compared to second constituents with more syllables (0.0 per cent to 4.3 per cent solid spellings). Hypothesis B8 can thus be confirmed if the 2 per cent lead on hyphenation for the first constituent is accepted: compounds containing one or more monosyllabic constituents favour solid spelling.

5.2.9 *Summary*

Section 5.2 investigates length-related variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Number of constituents of the compound
- Number of letters of the compound
- Number of letters of the individual constituents
- Ratio between the lengths of the constituents (letters)
- Compounds whose first two constituents are of very different length (letters)
- Number of syllables of the compound
- Number of syllables of the individual constituents

Table 5.19 *Grouped length of second constituent (in syllables) [Syll_2_r] and spelling in OHS_600*

			Length of second constituent (syllables)			Total
			1	2	3–5	
OHS	o	Count	112	70	18	200
		Expected Count	144.0	47.0	9.0	200.0
		% within Syll_2_r	25.9%	49.6%	66.7%	33.3%
	h	Count	126	65	9	200
		Expected Count	144.0	47.0	9.0	200.0
		% within Syll_2_r	29.2%	46.1%	33.3%	33.3%
	s	Count	194	6	0	200
		Expected Count	144.0	47.0	9.0	200.0
		% within Syll_2_r	44.9%	4.3%	0.0%	33.3%
Total	Count		432	141	27	600
	Expected Count		432.0	141.0	27.0	600.0
	% within Syll_2_r		100.0%	100.0%	100.0%	100.0%

- Ratio between the lengths of the constituents (syllables)
- Compounds whose constituents are of very different length (syllables).

Statistical testing revealed a strong effect of several independent variables on the dependent variable ‘compound spelling’ [OHS], as expected:

- A length of four or more syllables disfavours solid spelling. [B2]
- A length of eleven or more letters disfavours solid spelling. [B3]
- A ratio exceeding 2:1/1:2 in the number of letters of the constituents disfavours solid spelling. [B5]
- A ratio exceeding 2:1/1:2 in the number of syllables of the constituents disfavours solid spelling. [B6]
- A second constituent comprising three letters favours solid spelling. [B7]
- A monosyllabic second constituent favours solid spelling. [B8]

One hypothesis was only partly confirmed:

- Compounds whose first constituent comprises one syllable do favour solid spelling – but the proportion of hyphenation is relatively close. [B8]

Another result contradicts the expectations:

- Compounds whose first constituent comprises three letters do not favour solid spelling. [B7]

Furthermore, one can observe the following types of influence, even if the results do not permit statistical testing due to the small number of cases in the database:

- Compounds containing three or more constituents disfavour solid spelling. [B1]
- Compounds containing one constituent consisting of a single letter disfavour solid spelling. [B4]

Since the hypotheses for length needed to define in advance what should be considered a short or long compound or constituent, some findings go beyond the expectations formulated in the hypotheses. A reformulation of the results which includes these observations therefore yields more precise results:

- We can reduce the number of syllables at which a dispreference for solid spelling begins to three (instead of the four suggested in Hypothesis B2).

In addition to the results for the hypotheses under consideration, several other findings emerged from the detailed analysis of the material:

- Compounds containing three or more syllables favour open spelling. [B2]
- Bisyllabic compounds favour solid spelling. [B2]
- Bisyllabic compounds disfavour open spelling. [B2]
- Compounds comprising four to eight letters favour solid spelling. [B3]
- Compounds comprising four to eight letters disfavour open spelling. [B3]
- A ratio exceeding 2:1/1:2 in the number of letters of the constituents favours hyphenation. [B5]
- A ratio exceeding 2:1/1:2 in the number of syllables of the constituents favours open spelling. [B6]
- A long first constituent (seven or more letters) favours open spelling. [B7]
- A short first constituent (two letters) favours hyphenation. [B7]
- A second constituent comprising two letters favours hyphenation. [B7]
- A second constituent comprising three to five letters favours solid spelling. [B7]

- A second constituent comprising six or more letters disfavors solid spelling. [B7]
- A monosyllabic first constituent disfavors open spelling. [B8]
- A first constituent comprising two or more syllables favors open spelling. [B8]
- A second constituent comprising two or more syllables disfavors solid spelling. [B8]
- A second constituent comprising three or more syllables favors open spelling. [B8]

Since length was measured in three different ways (number of constituents, syllables and letters), it is interesting to compare the results. For all three length measures, the central result – that long compounds disfavor solid spelling – is identical and backed statistically where enough compounds are present in the dataset. In normal reading, parafoveal preview to the right of the currently fixated segment provides information on word length (Juhasz, Inhoff and Rayner 2005: 295). One may assume that experienced readers of English have formed the assumption that very long solid sequences of letters are either compounds or affixations and need to be analysed into constituents. Since affixations usually require solid spelling, the avoidance of solid spelling in very long compounds would have the advantage of permitting a clearer distinction from affixations before actually reading a sequence.

The results for the shortest compounds in terms of both syllables (two) and letters (four to eight) are also identical: as expected, short compounds favour solid spelling and tend to avoid open spelling. By contrast, the results are less clear for compounds with constituents of very different length: if the number of letters is considered, hyphenation is favoured, but if the number of syllables is used as the basis, there is a tendency towards open spelling. Of the fourteen overlapping compounds whose constituent length differs considerably regarding both letters and syllables, ten are unanimously spelled open in the dictionaries (e.g. *correspondence course*), whereas four are always hyphenated (e.g. *all-important*). This seems to suggest that in case of doubt, the number of syllables has priority over the number of letters as a determinant of compound spelling – an assumption supported by the important role of the number of syllables in the CompSpell spelling algorithm (cf. 6.2).

5.3 Frequency

According to Hasher and Zacks (1984), information on the frequency of events – such as the occurrence of particular words – is automatically encoded in the mind, and psycholinguistic studies commonly consider frequency an important variable for language processing. High frequency of occurrence potentially signals that a construction has unit status (Schönefeld 2006: 340) and eventually leads to its processing as a single unit (Bybee 2010: 96). The following sections investigate to what extent the frequency of compounds and their constituents and the frequency difference between the constituents play a role in the spelling of English compounds.

5.3.1 *Hypothesis C₁ – High-Frequency Compound*

High-frequency compounds favour solid spelling. → Refuted.

As far as the frequency of the whole compound is concerned, there is a reversal of expectations compared to other phenomena: while constructions with high frequency tend to conserve unusual features (cf. Aitchison 1991: 78), particularly in orthography, because “the eye will prefer what it is accustomed to see” (Fowler 1921: 12), the common expectation for compounds is to develop from open via hyphenated to solid spelling, especially if they are very frequent. This expectation is in line with Haspelmath’s (2008: 5) economy principle (“The more frequent a sign is, the shorter it is”) and an analogy to the observation that frequent repetition in spoken language leads to reduced pronunciation (Bybee 2003: 8–9, 58). It can be formulated as Hypothesis C₁:

C₁: High-frequency compounds favour solid spelling.

In order to test Hypothesis C₁, the frequency of each compound was determined by extracting all open, hyphenated and solid unlemmatised spelling variants from the written component of the British National Corpus (BNCwritten) by means of a Perl script which does not search across sentence boundaries in contrast to BNCweb and then adding the three resulting frequency values together. BNCwritten was used for the frequency data, since this balanced corpus of approximately 90 million words (cf. Aston and Burnard 1998: 28–29) is comparatively large, but still restricted to written language. The cut-off point for compounds with high frequency was determined by analogy to the approach adopted for the length-related hypotheses, so that at least 25 per cent of the

Table 5.20 *Grouped unlemmatised spelling-insensitive frequency in BNCwritten [Total_BNC_r] for the OHS_600 compounds*

Frequency		No of compounds	Per cent	Cumulative Per cent
0–17	low	153	25.5	25.5
18–126	mid	295	49.2	74.7
127–55,000	high	152	25.3	100.0
Total		600	100.0	

highest-frequency compounds formed one group. Based on the cumulative frequencies provided by SPSS, this corresponds to a frequency of 127 or higher. In order to meet the requirements for statistical testing, an analogical recoding was carried out for the roughly 25 per cent of compounds with the lowest frequency (including twelve OHS_600 compounds like *chick+flick*, *rocket+science* and *slumber+party* with no corpus hits) and for the intermediate segment. The median is at a frequency of forty-eight.

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped compound frequency' [Total_BNC_r] revealed a highly significant effect ($p = 0.000$). While the compounds with intermediate frequency show an almost perfect unmarked proportion of about one-third for each spelling variant, the low-frequency compounds display a clear preference for open spelling (45.8 per cent). As expected, high-frequency compounds are spelled solid more frequently than in a chance distribution (with 61 instead of 50.7 expected items and a proportion of 40.1 per cent instead of randomly distributed 33.3 per cent), but there are even slightly more hyphenations (sixty-two items = 40.8 per cent). As a consequence, Hypothesis C1 must be refuted: high-frequency compounds do not favour solid spelling.

The refutation of Hypothesis C1 is highly unexpected in view of the unanimous preference for solid spelling of high-frequency compounds observed in previous research (e.g. Sepp 2006: 87; Kuperman and Bertram 2013: 946). The results of the present study also contradict the additional dispreference for hyphenation with increasing frequency found by Kuperman and Bertram (2013: 946). While this could be due to the inclusion of too many compounds in the high-frequency category, the modification of the category boundaries does not change the present study's findings: even if the highest-frequency segment were reduced, hyphenations would still predominate (100 per cent in the top three,

Table 5.21 *Grouped unlemmatised spelling-insensitive frequency in BNCwritten [Total_BNC_r] and spelling in OHS_600*

			Frequency in BNCwritten			Total
			0–17 (low)	18–126 (mid)	127–55,000 (high)	
OHS	o	Count	70	101	29	200
		Expected Count	51.0	98.3	50.7	200.0
		% within Total_BNC_r	45.8%	34.2%	19.1%	33.3%
	h	Count	45	93	62	200
		Expected Count	51.0	98.3	50.7	200.0
		% within Total_BNC_r	29.4%	31.5%	40.8%	33.3%
	s	Count	38	101	61	200
		Expected Count	51.0	98.3	50.7	200.0
		% within Total_BNC_r	24.8%	34.2%	40.1%	33.3%
Total	Count		153	295	152	600
	Expected Count		153.0	295.0	152.0	600.0
	% within Total_BNC_r		100.0%	100.0%	100.0%	100.0%

50 per cent in the top ten and 55 per cent in the top twenty compounds).⁸ Instead, the differences in the results can be explained by differences in the approaches: since previous studies have tended to consider only nominal noun+noun compounds, whereas the present research studies compounding more generally, a separate analysis was carried out for the 289 n+n=n compounds in OHS_600. Table 5.22 demonstrates that this changes the picture completely: if only n+n=n compounds are considered, we do indeed find the clear predominance of solid spelling and the complete lack of hyphenation for the highest-frequency segment reported in the literature. This would seem to suggest that part of speech is an influential factor that has been unduly neglected in previous research.

5.3.2 *Hypothesis C2 – Low-Frequency Constituents*

Compounds containing two low-frequency constituents favour open spelling. → Refuted.

⁸ Note, however, that some high-frequency phrases skewed the corpus-based compound frequencies: thus the 55,000 hits in BNCwritten for the uniquely hyphenated OHS_600 noun compound *has-been* ('a person who used to be important or popular but has now been forgotten'; cf. LDOCE) are due to the fact that CompSpell cannot distinguish the compound with open spelling from the extremely frequent competing verb phrase *has been*.

Table 5.22 *Grouped unlemmatised spelling-insensitive frequency in BNCwritten [Total_BNC_r] and spelling of the noun+noun=noun compounds in OHS_600*

			Frequency in BNCwritten			Total
			0–17 (low)	18–126 (mid)	127–55,000 (high)	
OHS	o	Count	56	69	14	139
		Expected Count	40.4	70.7	27.9	139.0
		% within Total_BNC_r	66.7%	46.9%	24.1%	48.1%
	h	Count	3	4	0	7
		Expected Count	2.0	3.6	1.4	7.0
		% within Total_BNC_r	3.6%	2.7%	0.0%	2.4%
	s	Count	25	74	44	143
		Expected Count	41.6	72.7	28.7	143.0
		% within Total_BNC_r	29.8%	50.3%	75.9%	49.5%
Total	Count		84	147	58	289
	Expected Count		84.0	147.0	58.0	289.0
	% within Total_BNC_r		100.0%	100.0%	100.0%	100.0%

Another way in which frequency may exert some influence on English compound spelling concerns the compounds' constituents. Thus compounds containing two low-frequency constituents might be more likely to use open spelling in order to facilitate segmentation. This can be formulated as Hypothesis C2:

C2: Compounds containing two low-frequency constituents favour open spelling.

In order to determine constituent frequency, a lemmatised frequency list with the corresponding parts of speech was extracted from the written component of the British National Corpus in the CQP Edition of BNCweb (bncweb.lancs.ac.uk, 26 April 2012).⁹ Based on the part of speech that had been coded manually for the OHS_600 compound constituents, the program CompSpell retrieved part-of-speech-specific frequencies for each constituent, e.g. 5,423 for the noun *guide* and 2,117 for the verb *guide*. The frequency of grammatical words (cf. 5.6.1) was calculated by adding the respective frequencies for formally identical adjectives, adverbs, articles,

⁹ Note that this frequency list contains no open compounds, so that the constituents of open compounds will have counted towards the lemmatised frequency results for these constituents, whereas the constituents of hyphenated and solid compounds will have increased the compound frequency count for the hyphenated and solid listed compounds.

Table 5.23 *Frequency ranges for the first and second constituents in OHS_600*

	Low frequency (l)	Intermediate frequency (m)	High frequency (h)
Constituent 1	0–1,516 = 25.0%	1,517–21,568 = 49.7%	21,569–2,360,010 = 25.3%
Constituent 2	0–832 = 25.0%	833–20,617 = 50.0%	20,618–3,530,089 = 25.0%

conjunctions, interjections, prepositions and pronouns. The constituent frequencies for adverbs and formally indistinguishable adjectives (e.g. *double* in the verb *double+check*) were also added. Since part of speech had only been coded for the constituents of OHS_600 compounds, CompSpell calculated the approximate constituent frequency for all other Master List compounds by disregarding part of speech in the lemmatised BNCwritten frequency list and by summing all frequencies of matching forms.

The cut-off point for constituents with low, intermediate and high frequency was determined by following the same principle as in Section 5.3.1: for each constituent, the frequency-ordered OHS_600 dataset was subdivided into three groups in such a way that the proportion of low-frequency and high-frequency constituents represented 25 per cent counting from the bottom and the top, respectively. Slight deviations from this number are due to the fact that the cut-off frequency may be shared by several compounds.

In order to test hypothesis C2, Pearson’s chi-square test was carried out for the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘combined frequency ranges of the first two constituents’ [Freq_ivs2] for OHS_600. The results reached a level of significance of $p = 0.000$. Nevertheless, Hypothesis C2 had to be refuted: compounds containing two low-frequency constituents (ll) do not favour open spelling. With a proportion of 35.3 per cent (= 12 instead of expected 11.3 hits), the amount of open spelling in this group corresponds almost precisely to the expectations, whereas there is a preference for hyphenation (41.2 per cent) and a dispreference for solid spelling (23.5 per cent).

5.3.3 *Hypothesis C3 – High-Frequency Constituents*

Compounds containing two high-frequency constituents favour solid spelling. → Refuted.

Table 5.24 Combined frequency ranges of the first and second constituents [Freq_ivs2] and spelling in OHS_600
(l = low, m = mid, h = high)

		Combined frequency ranges of first and second constituents									
		Hh	hl	hm	lh	ll	lm	mh	ml	mm	
OHS	Count	12	4	18	14	12	43	10	16	71	
	Expected Count	17.3	13.3	20.0	12.3	11.3	26.3	20.3	25.3	53.7	
	% within Freq_ivs2	23.1%	10.0%	30.0%	37.8%	35.3%	54.4%	16.4%	21.1%	44.1%	
h	Count	27	32	24	17	14	6	13	48	19	
	Expected Count	17.3	13.3	20.0	12.3	11.3	26.3	20.3	25.3	53.7	
	% within Freq_ivs2	51.9%	80.0%	40.0%	45.9%	41.2%	7.6%	21.3%	63.2%	11.8%	
s	Count	13	4	18	6	8	30	38	12	71	
	Expected Count	17.3	13.3	20.0	12.3	11.3	26.3	20.3	25.3	53.7	
	% within Freq_ivs2	25.0%	10.0%	30.0%	16.2%	23.5%	38.0%	62.3%	15.8%	44.1%	
Total		52	40	60	37	34	79	61	76	161	

Since high-frequency constituents should be recognised and processed more quickly than others, the segmentation of compounds with two high-frequency constituents might be less dependent on physical cues at the constituent boundaries. At the same time, the use of solid spelling in such compounds would ensure that they are immediately perceived as a unit. This translates into the following hypothesis:

C3: Compounds containing two high-frequency constituents favour solid spelling.

The dataset and coding procedure used in the testing of Hypothesis C3 are identical to those used in Section 5.3.2. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'combined frequency ranges of the first two constituents' [Freq_1vs2] was highly significant ($p = 0.000$) for the OHS_600 compounds. Nonetheless, Hypothesis C3 has to be refuted: compounds combining two high-frequency constituents do not favour solid spelling (with 25.0 per cent instead of the expected 33.3 per cent). Instead, hyphenation clearly dominates in this group with a proportion of 51.9 per cent.

5.3.4 *Hypothesis C4 – Frequency Difference between Constituents*

A large difference in the frequencies of the constituents disfavours solid spelling. → Confirmed.

One general expectation of the present study is that spellers avoid concatenating heterogeneous constituents (cf. 5.12.2), and one way in which constituents may differ concerns their frequency, which yields the following specific hypothesis:

C4: A large difference in the frequencies of the constituents disfavours solid spelling.

Hypothesis C4 was tested by using the frequency data compiled for the testing of Hypothesis C2. At first, it was attempted to define a large difference in the frequencies of the constituents as a ratio exceeding 2:1/1:2, i.e. by using the same conventions as for length (cf. 5.2.5), but important differences emerged between the two variables: while a constituent cannot have a length of zero letters or syllables, it can have a frequency of zero hits in a corpus.¹⁰ Furthermore, differences in the length

¹⁰ Since divisions in which the second constituent is zero are mathematically impossible, all compounds comprising a constituent with a frequency of zero were coded with zero and disregarded.

of compound constituents are unlikely to reach similar extremes as those for frequency differences, since the range of possible values is larger for frequency data (particularly if derived from a large corpus). To reach a comparable ratio, compounds would necessitate constituents of unlikely length (e.g. 1,000 letters). Since a constituent frequency ratio exceeding 2:1/1:2 constituted the norm rather than the exception (with 450 instances among the OHS_600 compounds), a more extreme distribution of ratios [Freq_diff_12_r] was considered with the frequency ranges

- 1:0–1:49 ('no large difference')
- 1:50–1:99 ('a large difference')
- 1:100–1:999 ('a very large difference')
- 1:1000 and larger ('an extremely large difference').

The directionality of the differences (i.e. whether the first or second constituent has the higher frequency) was not considered.

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped frequency difference between the constituents' [Freq_diff_12_r] yielded highly significant ($p = 0.000$) results for the OHS_600 compounds. The larger the frequency difference between the constituents, the stronger the preference for hyphenation, with an increase from 22.3 per cent to 88.2 per cent. This is accompanied by an almost parallel decrease in both open and solid spellings, which go down to only one hit each in the most extreme set of compounds. From this we can conclude that Hypothesis C4 is supported by the data: a large difference in the frequencies of the constituents disfavours solid spelling.

Having considered the frequencies of the first and second constituents in relation to each other, their effect on the spelling of English compounds was also investigated separately in Hypotheses C4A and C4B:

C4A: The frequency of the first constituent has an effect on the spelling of biconstituent English compounds.

C4B: The frequency of the second constituent has an effect on the spelling of biconstituent English compounds.

The zero column is a special case: since the frequency of the constituents was determined by using a part-of-speech-sensitive lemmatised frequency list (cf. 5.3.2), the constituents with a frequency of zero are usually inflected forms, such as plural *knees* (in *knees+up*), genitive *banker's* (in *banker's+order*) or participles such as *handed* in *even+handed*. Compounds with one constituent with a frequency of zero are nearly always hyphenated (e.g. *habit-forming*) and practically never spelled solid (with the exception of *inbred*).

Table 5.25 *Grouped frequency difference between the constituents [Freq_diff_12_r] and spelling in OHS_600*

		Range of frequency differences between constituents						Total
		0	1: 1–49	1: 50–99	1: 100–999	1: 1000–38807		
OHS	o	Count	173	7	10	1	1	200
		Expected Count	151.0	7.0	21.0	5.7		200.0
		% within Freq_diff_12_r	38.2%	33.3%	15.9%	5.9%		33.3%
	h	Count	101	10	40	15	15	200
		Expected Count	151.0	7.0	21.0	5.7		200.0
		% within Freq_diff_12_r	22.3%	47.6%	63.5%	88.2%		33.3%
	s	Count	179	4	13	1	1	200
		Expected Count	151.0	7.0	21.0	5.7		200.0
		% within Freq_diff_12_r	39.5%	19.0%	20.6%	5.9%		33.3%
	Total	Count	453	21	63	17		600

Table 5.26 *Grouped frequency range of the first constituent [Freq_1_r] and spelling in OHS_600*

			Frequency of first constituent			Total
			low	mid	high	
OHS	o	Count	69	97	34	200
		Expected Count	50.0	99.3	50.7	200.0
		% within Freq_1_r	46.0%	32.6%	22.4%	33.3%
	h	Count	37	80	83	200
		Expected Count	50.0	99.3	50.7	200.0
		% within Freq_1_r	24.7%	26.8%	54.6%	33.3%
	s	Count	44	121	35	200
		Expected Count	50.0	99.3	50.7	200.0
		% within Freq_1_r	29.3%	40.6%	23.0%	33.3%
Total	Count		150	298	152	600
	Expected Count		150.0	298.0	152.0	600.0
	% within Freq_1_r		100.0%	100.0%	100.0%	100.0%

The categorisation into low-, mid- and high-frequency segments, which was necessary to meet the requirements for statistical testing, follows the principles in Section 5.3.2. Pearson's chi-square tests for the dependent variable 'compound spelling' [OHS] and the independent variables 'grouped frequency range of the first constituent' [Freq_1_r] and 'grouped frequency range of the second constituent' [Freq_2_r] were both found to be significant for the OHS_600 compounds, with $p = 0.000$ in both cases. Hypotheses C4A and C4B can thus be confirmed: independently from the frequency relation between the constituents, the frequencies of the first and second constituents both have a significant effect on the spelling of English compounds.

More specifically, a low-frequency first constituent favours open spelling (46.0 per cent), and with increasing frequency of the first constituent, there is a clear tendency to use more hyphenations (up to 54.6 per cent).

With regard to the second constituent, low frequency clearly favours hyphenation (62.7 per cent), whereas high frequency clearly disfavours open spelling (24.0 per cent) and correlates with an equal proportion of hyphenation and solid spelling (38.0 per cent).

5.3.5 Hypothesis C5 – Variation in Low-Frequency Compounds

Infrequent compounds vary more in their spelling than high-frequency compounds. → Confirmed.

Table 5.27 *Grouped frequency range of the second constituent [Freq_2_r] and spelling in OHS_600*

			Frequency of second constituent			Total
			low	mid	high	
OHS	o	Count	32	132	36	200
		Expected Count	50.0	100.0	50.0	200.0
		% within Freq_2_r	21.3%	44.0%	24.0%	33.3%
	h	Count	94	49	57	200
		Expected Count	50.0	100.0	50.0	200.0
		% within Freq_2_r	62.7%	16.3%	38.0%	33.3%
	s	Count	24	119	57	200
		Expected Count	50.0	100.0	50.0	200.0
		% within Freq_2_r	16.0%	39.7%	38.0%	33.3%
Total	Count		150	300	150	600
	Expected Count		150.0	300.0	150.0	600.0
	% within Freq_2_r		100.0%	100.0%	100.0%	100.0%

One may assume that the probability for standardisation processes to set in is considerably higher for high-frequency compounds than for low-frequency compounds, since the number of events during which changes might occur is decreased. As a consequence, one may also expect the spelling of infrequent words to display more variation than that of high-frequency words. This can be formulated as hypothesis C5:

C5: Infrequent compounds vary more in their spelling than high-frequency compounds.

Furthermore, once a tendency towards a particular spelling has emerged and one spelling variant has become very frequent, language users may feel no reason to use other variants anymore: these are prevented by analogy to the phenomenon of *blocking* in morphology, in which a form such as *stealer* is not used because of its well-established (i.e. frequent) alternative, *thief* (Aronoff 1976: 43; Burgschmidt 1973: 124).

In order to test Hypothesis C5, the whole Master List was used, because this permitted the consideration of a maximally large frequency range. The roughly 10,000 compounds were divided into stretches of about 2,000 compounds each, with increasing frequency for all spelling variants in BNCwritten [Total_BNC]. Subsequently, it was determined how many biconstituent compounds with unique spelling in the dictionaries (i.e. OHS_600 and OHS_extra compounds) occurred in each frequency

Table 5.28 *Corpus frequency and spelling variation in the dictionaries for the Master List compounds*

Frequency range for all compound spellings in BNCwritten	Number of biconstituent compounds	Number of biconstituent compounds with unanimous spelling in all dictionaries (OHS_600 + OHS_extra)	Number of biconstituent compounds with spelling variation in the dictionaries	Proportion of biconstituent compounds with spelling variation
very low (0–5)	1,948	45+396 = 441	1,507	77.4%
low (6–17)	2,006	112+775 = 887	1,119	55.8%
intermediate (18–40)	1,818	126+837 = 963	855	47.0%
high (41–118)	1,866	156+977 = 1,133	733	39.3%
very high (119–193,754)	1,620	161+879 = 1,040	580	35.8%
TOTAL	9,258	4,464	4,794	51.8%

band. The number of compounds with spelling variation was then calculated by subtracting that number from the size of the frequency band, and the proportion of spelling variation was calculated separately for each frequency band. Variation was determined based on the dictionaries to make the results more comparable to the remainder of the study and to avoid mistakes based on the incorrect classification of phrasal sequences as open compounds in unedited corpus data (cf. 5.3.2).

It is immediately obvious from the last column in Table 5.28 that variation in the dictionaries increases with decreasing corpus frequency. This result was confirmed by a chi-square test in *Excel* ($p = 0.000$). Hypothesis C5 is therefore supported by the data: low-frequency compounds vary more in their spelling than high-frequency compounds.

5.3.6 Summary

Section 5.3 investigates frequency-related variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Frequency of the compound with corresponding part of speech in the six dictionary lemma lists (e.g. LDOCE, MED, . . .)
- Frequency of the whole compound in BNCwritten
- Frequency of the individual compound constituents in the lemmatised BNCwritten frequency list
- Ratio between the frequencies of the constituents in the lemmatised BNCwritten frequency list
- Compounds whose first two constituents are of very different frequency.

Statistical testing revealed a strong effect of several independent variables on the dependent variable 'compound spelling' [OHS] for the OHS_600 compounds. The following hypotheses were confirmed:

- A large difference in the frequencies of the constituents disfavours solid spelling. [C4]
- Low-frequency compounds vary more in their spelling than high-frequency compounds. [C5]

In addition, the results for [C4] permitted an elaboration regarding precision:

- A large difference in the frequencies of the constituents (i.e. a ratio exceeding 1:50/50:1) favours hyphenation. [C4]

By contrast, the findings for three hypotheses contradict the expectations:

- High-frequency compounds do not favour solid spelling. [C1]
- Two low-frequency constituents do not favour open spelling. [C2]
- Two high-frequency constituents do not favour solid spelling. [C3]

In addition to the results for the hypotheses under consideration, several other findings emerged from the detailed analysis of the material:

- High-frequency compounds disfavour open spelling. [C1]
- Low-frequency compounds favour open spelling. [C1]
- Two low-frequency constituents favour hyphenation. [C2]
- Two low-frequency constituents disfavour solid spelling. [C2]
- Two high-frequency constituents favour hyphenation. [C3]
- A low-frequency first constituent favours open spelling. [C4A]
- A high-frequency first constituent favours hyphenation. [C4A]
- A low-frequency second constituent favours hyphenation. [C4B]
- A high-frequency second constituent disfavors open spelling. [C4B]

5.4 Phonology

The following sections investigate which phonological factors might influence variant selection in English compound spelling.

5.4.1 Hypothesis D1 – Silent <e> Preceding the Constituent Joint

Compounds with a silent <e> preceding the constituent joint (e.g. *rope+ladder*) disfavour solid spelling. → Refuted.

Word-final silent <e> in the spelling of English words determines the phonological quality of the preceding orthographic vowel and makes it “say its name” (Okada 2005: 175): thus the grapheme <a> in *bat* is pronounced /æ/, whereas the same character in *hate* corresponds to the diphthong /eɪ/. Since the phonological quality of vowel graphemes such as <a> in the example *hate* can thus only be determined in hindsight, after having processed the word-final silent <e>, it may be particularly important to have a visually salient constituent boundary following this grapheme in English compounds. As a consequence, one may expect the avoidance of concatenation after a constituent-final silent <e>:

D1: Compounds with a silent <e> preceding the constituent joint (e.g. *rope+ladder*) disfavour solid spelling.

In order to test this hypothesis, all sequences of <e> followed by a plus sign signalling a constituent boundary were coded with *e* in a separate column of the database for the OHS_600 compounds (e.g. *store+house*), and it was checked whether the omission of the <e> would result in a different vowel quality of the preceding vowel. Since the unmarked pronunciation of a vowel grapheme is not always easy to determine, particularly when it is combined with certain consonants, it was decided to consider only those instances with constituent-final <e> where the preceding single vowel was pronounced like its isolated alphabetic variant (e.g. *rope+ladder*). These were coded by replacing the code *e* with *s*. Double-vowel graphemes (e.g. *feeble*, *double*) were not considered, because the extent to which their pronunciation is influenced by silent <e> is difficult to determine.

One hundred twenty-four of the OHS_600 compounds contain a silent <e> at the end of the first constituent, which corresponds to the surprisingly large proportion of 20.7 per cent. Pearson’s chi-square test for the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘silent <e> at the end of the first constituent’ [Silent_e] for

Table 5.29 *Possible combinations of stress and spelling in compounds based on the examples in Bauer (1983: 104 and 1998: 79)*

Spelling	Fore-stress	Back stress
OPEN	<i>'trade name</i>	<i>,bank 'holiday</i>
HYPHENATED	<i>'strip-show</i>	<i>,trade-'union</i>
SOLID	<i>'bankrate</i>	<i>,man'kind</i>

the OHS_600 compounds yielded no significant results ($p = 0.687$). Since the number of cases is large enough to permit statistically valid conclusions, Hypothesis D1 must be refuted: silent <e> at the end of the first constituent has no consequence on the spelling of English compounds, regardless of whether it results in a change of vowel quality. This finding is relatively surprising, particularly for those cases where the presence of the <e> retroactively determines the quality of the preceding vowel.

5.4.2 *Hypothesis D2 – Main Stress on First Constituent*

Compounds with main stress on the first constituent (e.g. *'gold+fish*) disfavour open spelling. → Refuted.

A supra-segmental feature frequently mentioned in relation to compounding is stress (cf. 2.1.1.2). While all conceivable combinations of stress and spelling variant are attested in the literature (cf. Table 5.29), Plag’s (2010) study on compound stress identified spelling as the strongest predictor for stress assignment.

One may, however, question whether this order truly reflects the underlying causality, i.e. whether stress really depends on spelling, or whether it is actually the other way round. Possibly both spelling and stress correlate with another factor as the true underlying reason. For example, word stress is frequently explained by drawing on semantics: Bauer (1998: 71) reports a preference for stress on the first constituent in compounds with the semantic relation ‘B used for A’ (*pruning shears*). Noun+noun compounds with a first constituent which is a proper noun (*Ilkley Moor*) or which designates a location (*kitchen sink*), time (*night watchman*) or material (*cotton dress*) are likely to be stressed on the second constituent, whereas a first constituent denoting a purpose or destination (*toothbrush*), an originator (*rainwater*) or a resemblance (*goldfish*) is invariably tied up

with fore-stress (Marchand 1960a: 15–17).¹¹ In view of the open and solid spelling of Marchand's examples, one may formulate the following hypothesis:

D2: Compounds with main stress on the first constituent (e.g. 'gold+fish) disfavour open spelling.

In order to test Hypothesis D2, CompSpell automatically coded stress in the Master List by using the electronic part-of-speech-specific list of compound stress patterns from the *Macmillan English Dictionary for Advanced Learners* (MED), which had kindly been extracted by the publisher. This list of (mainly) open and hyphenated compounds was used in full awareness of the fact that dictionaries may differ with regard to stress assignment (cf. Plag 2010: 249). However, it is commonly observed in the literature that compounds are not assigned stress consistently anyway – neither by the speech community in general nor by individual speakers (Bauer 1983: 103; Hacken 1994: 34; Sepp 2006: 95–96). In order to determine how much the selection of the MED as a source influences the results, the stress of fifty random compounds was compared to the stress patterns in the *Longman Pronunciation Dictionary* (LPD 2008). While fourteen compounds were not contained in the LPD (which gives additional support to the use of a general-language dictionary for the coding of stress patterns), both dictionaries agreed in the overwhelming majority of thirty-three cases for the remaining thirty-six compounds (at least in one of the LPD's listed variants), and merely three items have primary stress in the MED but secondary stress in the LPD (*no+frills*, *on+screen* and *blood+brother*).

The MED list contains word stress information for 5,273 of the Master List compounds. Main stress is indicated by a preceding <~>, secondary stress by a preceding <@>, e.g. @able -seaman. This information was reduced to the sequence of stress marks @~ in a separate column of the database. Three hundred twenty-four of the OHS_600 compounds were coded in this way. Of the remaining cases, 272 had to be coded manually based on the MED, and four compounds for which no stress information

¹¹ According to Marchand (1960a: 18),

"Many forestressed compounds denote an intimate, permanent relationship between the two significates to the extent that the compound is no longer to be understood as the sum of the constituent elements. A summer-house, for instance, is not merely a house inhabited in summer, but a house of a particular style and construction which make it suitable for the warm season only. Two-stressed combinations of the type *stone wall* never have this character. A syntactic group is always analysable as the additive sum of its constituents."

could be retrieved in that dictionary were coded following the LPD (*weather+man*, *bathing+costume*, *high+risk*, *water+repellent*). The stress patterns were then recoded with <1> for stress on the first constituent, <2> for stress on the second constituent and <3> for the small number of unusual stress patterns in the Master List (e.g. @@@~ or @~@) in a separate column.

Four hundred and five (i.e. about two-thirds) of the OHS_600 compounds have main stress on the first constituent, as one might have expected from such established lexemes. It is interesting to note, however, that another third (195 compounds) of the OHS_600 compounds does not meet this expectation and that so-called *phrasal stress* is prevalent even in this particular group of compounds. If we compare this finding to the stress patterns of the compounds contained in up to four dictionaries (Master_1–4), we find that merely 55.0 per cent of the 1,948 compounds with stress indications in the MED are stressed on the first constituent as against 42.8 per cent on the second and 2.2 per cent on the third. This shift towards phrasal stress might be taken as an indication that the present study confirms the role of stress on the first constituent as an indicator of lexicalisation (cf. Plag, Kunter and Lappe 2007).

In order to test Hypothesis D2, Pearson's chi-square test was conducted for OHS_600 with the dependent variable 'compound spelling' [OHS] and the independent variable 'stress pattern' [Stress]. The results are highly significant ($p = 0.000$). Compounds with so-called *compound stress* on the first constituent clearly favour solid spelling (48.4 per cent) and disfavour hyphenation (21.5 per cent). Since the proportion of 30.1 per cent open spellings in this group is very close to the expected average of 33.3 per cent, however, Hypothesis D2 needs to be refuted: compounds with main stress on the first constituent do not disfavour open spelling.

5.4.3 Hypothesis D3 – Main Stress on Second Constituent

Compounds with main stress on the second constituent (e.g. *apple+pie*) disfavour solid spelling. → Confirmed.

Since stress on the second constituent of biconstituent compounds is often referred to as *phrasal stress* and in view of the fact that phrases use open spelling as a rule, the following expectation can be formulated:

D3: Compounds with main stress on the second constituent (e.g. *apple+pie*) disfavour solid spelling.

Table 5.30 *Main stress [Stress] and spelling in OHS_600*

			Main stress on		Total
			1st constituent	2nd constituent	
OHS	o	Count	122	78	200
		Expected Count	135.0	65.0	200.0
		% within Stress	30.1%	40.0%	33.3%
	h	Count	87	113	200
		Expected Count	135.0	65.0	200.0
		% within Stress	21.5%	57.9%	33.3%
	s	Count	196	4	200
		Expected Count	135.0	65.0	200.0
		% within Stress	48.4%	2.1%	33.3%
Total	Count		405	195	600
	Expected Count		405.0	195.0	600.0
	% within Stress		100.0%	100.0%	100.0%

Hypothesis D3 was tested simultaneously with Hypothesis D2 (cf. Table 5.30). Stress pattern [Stress] emerged as a highly significant variable ($p = 0.000$) in Pearson's chi-square test with the dependent variable 'compound spelling' [OHS] and the independent variable 'stress pattern' [Stress] for OHS_600. The proportion of only 2.1 per cent solid spellings among the back-stressed compounds clearly supports Hypothesis D3: compounds with main stress on the second constituent disfavour solid spelling. While this results in the expected preference for open spelling (40.0 per cent) that one would expect from the terminological link to phrases, the tendency towards hyphenation is, however, even stronger in this group (57.9 per cent). This finding is consequently in line with the correlation between back stress and hyphenation suggested by older style guides such as Hart (1957: 35) or Morton Ball (1951: 9). The data also support Plag's (2010: 265) observation that solid compounds tend to be stressed on the first constituent, whereas open compounds are "much more variable in their stress pattern": in the present study, solid compounds are overwhelmingly fore-stressed (98.0 per cent), whereas open compounds have a less clear ratio of 61:39 for stress on the first and second constituents, respectively.

5.4.4 Hypothesis D4 – Noun with Main Stress on Second Constituent

Noun compounds with main stress on the second constituent (e.g. *gas+cooker*) favour open spelling. → Confirmed.

That part of speech may play an important role in combination with stress can be inferred from Quirk et al.'s (1985: 1570–1578) account of compounding (cf. 5.6.3): all subcategories with phrasal stress in their example words (*hard-working, quick-frozen, grass-green, grey-green*) tend towards hyphenation and are exemplified with adjectives. Phrasal stress is also possible in some nominal example words of the usually fore-stressed types '*wind, mill*' (e.g. back-stressed *gas 'cooker*) and '*dark, room*' (e.g. back-stressed *fancy 'dress*), which are then spelled open. Phrasal stress may thus result in hyphenation or open spelling, depending on the part of speech. This was tested in Hypotheses D4 and D5. Hypothesis D3, which assumes that main stress on the second constituent leads to the avoidance of solid spelling, was thus refined regarding part of speech:

D4: Noun compounds with main stress on the second constituent (e.g. *gas+ 'cooker*) favour open spelling.

In order to test Hypothesis D4, a separate file containing only the 426 noun compounds from OHS_600 was created and subjected to Pearson's chi-square test, with the dependent variable 'compound spelling' [OHS] and the independent variable 'stress pattern' [Stress]. The results are highly significant ($p = 0.000$). This is due to the extremely large proportion of open noun compounds with main stress on the second constituent, which is almost twice as high as expected (78 instead of 39.9 hits). Hypothesis D4 can therefore be confirmed: noun compounds with main stress on the second constituent favour open spelling.

Table 5.31 *Main stress [Stress] and spelling for the noun compounds in OHS_600*

			Stress		Total
			1st constituent	2nd constituent	
OHS	o	Count	122	78	200
		Expected Count	160.1	39.9	200.0
		% within Stress	35.8%	91.8%	46.9%
	h	Count	39	6	45
		Expected Count	36.0	9.0	45.0
		% within Stress	11.4%	7.1%	10.6%
	s	Count	180	1	181
		Expected Count	144.9	36.1	181.0
		% within Stress	52.8%	1.2%	42.5%
Total	Count		341	85	426
	Expected Count		341.0	85.0	426.0
	% within Stress		100.0%	100.0%	100.0%

5.4.5 Hypothesis D5 – Adjective with Main Stress on Second Constituent

Adjective compounds with main stress on the second constituent (e.g. *hard+* 'working') favour hyphenation. → Tentatively confirmed.

In Quirk et al.'s (1985: 1570–1578) account of compounding (cf. 5.6.3), all subcategories with phrasal stress in the example words (*hard-working*, *quick-frozen*, *grass-green*, *grey-green*) use adjective compounds as examples and tend towards hyphenation. Hypothesis D3, which assumes that main stress on the second constituent leads to the avoidance of solid spelling, can thus be refined regarding part of speech:

D5: Adjective compounds with main stress on the second constituent (e.g. *hard+* 'working') favour hyphenation.

In order to test Hypothesis D5, a separate file containing only the 157 adjective compounds from OHS_600 was created and subjected to Pearson's chi-square test, with the dependent variable 'compound spelling' [OHS] and the independent variable 'stress pattern' [Stress]. Since the one cell with an expected count below five corresponds to a proportion of 25 per cent, the otherwise highly significant result ($p = 0.000$) is not statistically valid.

Nonetheless, adjective compounds with main stress on the second constituent clearly favour hyphenation (with 102 instead of 95.4 hits). Hypothesis D5 can thus be tentatively confirmed: adjective compounds with main stress on the second constituent favour hyphenation – but so do adjectives in general (with a proportion of 79.2 per cent in the fore-stressed category). Since open spelling is completely avoided in the OHS_600

Table 5.32 Main stress [Stress] and spelling for the adjective compounds in OHS_600

			Stress		Total
			1st constituent	2nd constituent	
OHS	h	Count	42	102	144
		Expected Count	48.6	95.4	144.0
		% within Stress	79.2%	98.1%	91.7%
	s	Count	11	2	13
		Expected Count	4.4	8.6	13.0
		% within Stress	20.8%	1.9%	8.3%
Total	Count		53	104	157
	Expected Count		53.0	104.0	157.0
	% within Stress		100.0%	100.0%	100.0%

adjectives, this suggests that part of speech is a powerful variable in compound spelling variant selection.

5.4.6 Summary

Section 5.4 investigates phonology-related variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Occurrence of silent <e> before the constituent joint (e.g. in *race+horse*)
- Stress pattern (main stress on first or second constituent).

Statistical testing revealed a strong effect of several independent variables on the dependent variable 'compound spelling' [OHS] for OHS_600 or subsamples. The following hypotheses were confirmed:

- Main stress on the second constituent disfavors solid spelling. [D3]
- Noun compounds with main stress on the second constituent favour open spelling. [D4]

The findings for some of the other hypotheses, by contrast, contradict the expectations:

- Silent <e> preceding the constituent joint does not disfavor solid spelling.
- Main stress on the first constituent does not disfavor open spelling. [D2]

The following result can only be considered as indicative of a tendency, since it is not backed by statistical validity:

- Adjective compounds with main stress on the second constituent prefer hyphenation – but so do adjectives in general. [D5]

In addition to the results for the hypotheses under consideration, several other findings emerged from the detailed analysis of the material:

- Main stress on the first constituent favours solid spelling. [D2]
- Main stress on the first constituent disfavors hyphenation. [D2]
- Main stress on the second constituent favours open and particularly hyphenated spelling. [D3]

Some of the additional findings concern combinations of variables:

- Noun compounds with main stress on the first constituent favour solid spelling. [D4]
- Noun compounds with main stress on the first constituent disfavour open spelling. [D4]
- Noun compounds with main stress on the second constituent disfavour solid spelling. [D4]

5.5 Morphology

Morphology as the next variable under consideration covers all properties related to the level of morphemes (free or bound, lexical or grammatical), including word formation.

5.5.1 Affixes

The following sections investigate the assumption that the presence of affixes influences the spelling of English compounds. Since lexical and grammatical affixes fulfil different functions, their influence was tested in separate hypotheses.

5.5.1.1 Hypothesis EI – Lexical Suffix Preceding the Constituent Joint

A lexical suffix preceding the constituent joint (e.g. *amusement+park*) disfavors solid spelling. → Confirmed.

Lexical suffixes may be followed by other lexical suffixes (e.g. *-al* in *nationality* by *-ity*) and particularly by inflectional suffixes (e.g. *-er* in *teachers* by *-s*), but they are prototypically located at the end of lexemes. As a consequence, lexical suffixes are usually followed by a blank or some punctuation mark. For that reason, their presence at the end of a non-final compound constituent may exert some influence on the choice of orthographic variant by increasing the tendency to use open spelling or hyphenation:

EI: A lexical suffix preceding the constituent joint (e.g. *amusement+park*) disfavors solid spelling.

In order to test this hypothesis, non-compound-final lexical suffixes were coded in the OHS_600 compound set, but only if they attached to a discernible base (e.g. *poetic+licence*). This excludes e.g. *legal* or *magic*, but minor changes in the form of the base were tolerated (e.g. *carri+age* vs. *carry*).

Table 5.33 *Non-compound-final lexical suffixes [Nonfin_lex_suff] in OHS_600*

Non-compound-final lexical suffix	Frequency	Examples
-age	2	carriage
-al	5	artificial
-en	2	golden
-ence	1	correspondence
-er	3	roller
-ery	1	cookery
-ic	3	poetic
-ice	2	service
-ing	19	swimming
-ive	1	negative
-ly	1	prickly
-ry	1	registry
-th	1	sixth
-tion	1	reception
-ty	6	safety
-y	3	baby

Table 5.33 gives an overview of the fifty-two non-final lexical suffixes contained within the OHS_600 compounds. In view of the relatively large number of types and the relatively small number of tokens, recoding was necessary to meet the requirements of statistical testing by treating all instances of non-final lexical suffixes jointly in a binary distinction from the unmarked cases [Nonfin_lex_suff_r].

Pearson’s chi-square test for the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘presence of a non-compound-final lexical suffix’ [Nonfin_lex_suff_r] was highly significant ($p = 0.000$) for OHS_600. None of the compounds with a lexical suffix preceding the constituent joint uses solid spelling. Instead, there is an overwhelming tendency towards open spelling in this group (90.4 per cent). Hypothesis E1 is thus supported by the data: a lexical suffix preceding the constituent joint disfavours solid spelling, presumably because the prototypical use of suffixes at the end of words makes them function as orthographic closing sequences in this position.

5.5.1.2 *Hypothesis E2 – Inflection Preceding the Constituent Joint*

An inflectional suffix preceding the constituent joint (e.g. *equals sign*) disfavours solid spelling. → Tentatively confirmed.

Table 5.34 *Grouped non-compound-final lexical suffixes [Nonfin_lex_suff_r] and spelling in OHS_600*

			Non-compound-final lexical suffix		Total
			-	+	
OHS	o	Count	153	47	200
		Expected Count	182.7	17.3	200.0
		% within Nonfin_lex_suff_r	27.9%	90.4%	33.3%
	h	Count	195	5	200
		Expected Count	182.7	17.3	200.0
		% within Nonfin_lex_suff_r	35.6%	9.6%	33.3%
	s	Count	200	0	200
		Expected Count	182.7	17.3	200.0
		% within Nonfin_lex_suff_r	36.5%	0.0%	33.3%
	Total	Count	548	52	600
		Expected Count	548.0	52.0	600.0
		% within Nonfin_lex_suff_r	100.0%	100.0%	100.0%

Even if English is a language with comparatively little inflection, the constituents of compounds may contain grammatical information in the form of inflectional suffixes, e.g. in genitive compounds such as *cat's+cradle* (cf. 5.1.3.1). Other inflectional morphemes occasionally occurring in compounds are plural {S} (e.g. in *games mistress* or *drugs courier*), comparative *-er* (e.g. in *lower+case*), superlative *-est* (e.g. in *lowest+common+denominator*) and third person singular {S} (e.g. in *equals sign* or phrase compounds such as the plant name *love-lies-bleeding*). Since inflectional suffixes usually occur word-finally before a space or a punctuation mark and do not permit other suffixes to follow them, one might expect the effect discussed in Hypothesis E1 for lexical suffixes (cf. 5.5.1.1) to be even stronger for inflectional suffixes:

E2: An inflectional suffix (e.g. plural, genitive, present/past participle, superlative) preceding the constituent joint disfavours solid spelling.

In order to test Hypothesis E2, all instances of regular and irregular inflection at the end of the first constituent [Nonfin_gr_suff] were coded for OHS_600 (cf. Table 5.35 for an overview with examples). Note that *-ing* and *-ed* could be interpreted as either lexical or grammatical suffixes or as situated on a gradient between lexical and grammatical

Table 5.35 *Non-compound-final grammatical suffixes [Nonfin_gr_suff] in OHS_600*

Code	Description	Frequency	Examples
s_pl	plural <i>s</i>	4	<i>systems+analyst</i>
s_gen	genitive <i>s</i>	2	<i>banker's+order</i>
s_sing	third person singular <i>s</i>	1	<i>has+been</i>
ing	present participle suffix	0	
n	past participle suffix	5	<i>jumped+up</i> <i>cleft+palate</i> <i>broken+down</i>
ed	past tense	0	
comp	comparative <i>-er</i>	0	
sup	superlative <i>-est</i>	0	
	TOTAL	12	

suffixes, depending on the specific instance of occurrence. Since the compound-medial use of *-ing* does not correspond to the literal meaning derived from a syntactic analysis of the compounds (which, for *walking stick*, would yield 'a stick that walks'), it was considered a lexical rather than a grammatical suffix in that context (yielding the more sensible analysis 'a stick for walking'; note the different stress pattern). The non-compound-final past participle, by contrast, was always considered grammatical inflection due to its semantics (e.g. in *cleft+palate*; cf. OED s.v. *-ed¹* and *-ed²*).

In view of the relatively large number of types and the relatively small number of tokens, recoding was necessary to meet the requirements of statistical testing, and all instances of non-compound-final inflection were treated jointly in a binary distinction from the unmarked cases [Nonfin_gr_suff_r]. Pearson's chi-square test was then carried out for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped non-compound-final grammatical suffix' [Nonfin_gr_suff_r] for OHS_600. While the results are significant ($p = 0.017$), they are not statistically valid due to the large number of expected counts below five. Nevertheless, Table 5.36 permits the observation of the same tendency as for the non-compound-final lexical suffixes: solid spelling is avoided completely in the marked group, whereas the unmarked cases display a completely unmarked distribution (with almost precisely 33.3 per cent for each spelling variant). Even if statistical backing is lacking, the results suggest that Hypothesis E2 can be tentatively confirmed: an inflectional suffix preceding the constituent joint disfavours solid spelling.

Table 5.36 *Grouped non-compound-final grammatical suffixes [Nonfin_gr_suff_r] and spelling in OHS_600*

			Non-compound-final grammatical suffix		Total
			–	+	
OHS	o	Count	196	4	200
		Expected Count	196.0	4.0	200.0
		% within Nonfin_gr_suff_r	33.3%	33.3%	33.3%
	h	Count	192	8	200
		Expected Count	196.0	4.0	200.0
		% within Nonfin_gr_suff_r	32.7%	66.7%	33.3%
	s	Count	200	0	200
		Expected Count	196.0	4.0	200.0
		% within Nonfin_gr_suff_r	34.0%	0.0%	33.3%
	Total	Count	588	12	600
		Expected Count	588.0	12.0	600.0
		% within Nonfin_gr_suff_r	100.0%	100.0%	100.0%

5.5.1.3 *Hypothesis E3 – Prefix Following the Constituent Joint*

A prefix following the constituent joint (e.g. in *sneak+preview*) disfavours solid spelling. → Tentatively confirmed.

The line of argument used for suffixes in Sections 5.5.1.1 and 5.5.1.2 can also be applied to prefixes with the necessary changes: in English, prefixes are typically preceded by a space if they do not occur in the beginning of a new line, page etc. One may therefore expect that compounds containing a prefix at the beginning of a non-initial constituent tend to use open spelling, since that variant places a space before the non-initial prefix:

E3: A prefix following the constituent joint (e.g. in *sneak+**preview***) disfavours solid spelling.

In order to test this hypothesis, the presence of prefixes following the constituent joint was coded manually for the OHS_600 compounds (e.g. *non-* in the second constituent of *no+**nonsense***), but only if they were followed by a free base that exists in present-day English. Consequently, neither *ad-* in *personal ad* nor *re-* in *water+repellent* were counted. The word-initial morpheme *a-*, which is sometimes historically a preposition, was always considered a prefixation (e.g. in *go+ahead*) and standardised in the coding in instances of minimal formal change (e.g. in *ill+assorted*).

Table 5.37 *Non-compound-initial prefixes [Nonini_pref]
in OHS_600*

Code	Frequency	Examples
<i>a-</i>	3	<i>go+ahead</i>
<i>ab-</i>	1	<i>child+abuse</i>
<i>com-</i>	1	<i>glove+compartment</i>
<i>in-</i>	1	<i>artificial+insemination</i>
<i>non-</i>	1	<i>no+nonsense</i>
<i>para-</i>	1	<i>golden+parachute</i>
<i>re-</i>	1	<i>day+return</i>

The OHS_600 list contains the non-compound-initial prefixes [Nonini_pref] listed in Table 5.37.

The small number of types and tokens prevents a statistically significant analysis of the nine compounds containing non-initial prefixes. However, the fact that six of these are spelled open, three hyphenated and not a single one spelled solid supports Hypothesis E3 and suggests that the presence of a prefix following the constituent joint disfavors solid spelling indeed. The reason is analogous to that for the lexical and grammatical suffixes: since prefixations are expected to occur in the beginning of a sequence, this may incite spellers to visually create such an initial position by using spacing or hyphenation rather than run-on letters. The present study’s result is also in line with (although not identical to) Rakić’s (2009: 62) finding that those noun+noun compounds containing a prefixation in any position “are almost always written open”.

5.5.1.4 *Hypothesis E4 – Compound-Final -ing/-ed/-er*

Compounds containing the suffixes *-ing*, *-ed* or *-er* in compound-final position favour hyphenation. → Confirmed.

While the previous sections focused on affixes preceding or following the constituent joint, several compound-final suffixes are also expected to have an effect on spelling: according to Partridge (1953: 142), British English uses an obligatory hyphen in “noun + agential noun from a transitive verb” (e.g. *engine-driver*), except in some long-established words (e.g. *bricklayer*). This suggests a special role of the suffix *-er* and its variants. Furthermore, Partridge (1953: 142) posits an obligatory hyphen in British English “[n]oun + gerund of a transitive verb” (e.g. *labour-saving*), which suggests that the suffix *-ing* may also have a characteristic effect. Since the suffix *-ing*

Table 5.38 Compound-final -ing, -ed, -er and variants [Final_ingeder] in OHS_600

Code	Description	Example	Frequency
ing	-ing	<i>earth-shattering</i>	24
ed	past participle	<i>double-edged</i>	54
ed_irr	irregular past participle	<i>deep-set</i>	17
ed_pt	past tense	<i>also-ran</i>	1
er	-er	<i>snake charmer</i>	8
er_or	-or	<i>safety razor</i>	2

is frequently discussed by comparison to the suffix *-ed* (e.g. in *clear+sighted*), it makes sense to extend the expected spelling behaviour to this suffix as well. This permits the formulation of one joint hypothesis for three types of compound:

E4a/b/c: Compounds containing the suffixes

- a) *-ing*
- b) *-ed*
- c) *-er*

in compound-final position favour hyphenation.

In order to test this hypothesis, compound-final *-ed*, *-ing* and *-er* were coded for the OHS_600 compounds when they had morphemic status (which was not the case in semantically empty sequences such as the <er> at the end of *wrapping+paper*). Variants such as the *-or* in *copy+editor* were also accepted. While *-ing* was coded identically in adjectival constituents (*peace+loving*) and nominal constituents (*race+meeting*), an additional distinction had to be made for *-ed*, since the past participle function was sometimes realised by an irregular form, e.g. in *clear+cut*.

In order to test Hypothesis E4, Pearson's chi-square test was carried out for the dependent variable 'compound spelling' [OHS] and the grouped independent variable 'presence of the compound-final suffixes *-ing*, *-ed* or *-er*' [Final_ingeder_r] in OHS_600. For this binary distinction, the test yielded highly significant results ($p = 0.000$), which can be attributed to the avoidance of solid spelling (9 per cent) and the predominance of hyphenations (85 per cent) in the marked compounds. Hypothesis E4 is therefore supported by the data: compounds containing the suffixes *-ing*, *-ed* or *-er* (or their variants) in compound-final position favour hyphenation. This can be partly explained by Sepp's (2006: 113) argument that a certain

Table 5.39 *Grouped compound-final -ing, -ed or -er [Final_ingeder_r] and spelling in OHS_600*

			Compound-final <i>-ing</i> , <i>-ed</i> or <i>-er</i>		Total
			-	+	
OHS	o	Count	190	10	200
		Expected Count	164.7	35.3	200.0
		% within Final_ingeder_r	38.5%	9.4%	33.3%
	h	Count	109	91	200
		Expected Count	164.7	35.3	200.0
		% within Final_ingeder_r	22.1%	85.8%	33.3%
	s	Count	195	5	200
		Expected Count	164.7	35.3	200.0
		% within Final_ingeder_r	39.5%	4.7%	33.3%
Total	Count		494	106	600
	Expected Count		494.0	106.0	600.0
	% within Final_ingeder_r		100.0%	100.0%	100.0%

amount of hyphenation in compounds containing *-er* or *-ing* may be attributable to a “reluctance to concatenate noncanonical lexical orderings”, such as the object-verb structure of compounds like *house-painter*. One could argue that readers reinterpret such structures by reversing the order of the constituents in the compound’s surface structure to arrive at the underlying deep structure (‘someone paints the house’). The orthographic analysis would thus support segmentation as a first step towards the reordering. If we consider a compound’s spelling as an indication of its morphological structure (cf. 5.5.2), yet another explanation for the avoidance of solid spelling in this group of compounds can be found, namely the easier distinction from suffixations of complex bases. However, since neither of these explanations can account for the avoidance of open spelling, the postulation of a hyphenation principle for marked compounds (in this case regarding the non-canonical lexical ordering) might make even more sense (cf. also 7.1).

5.5.1.5 *Hypothesis E5 – Constituent-Internal Hyphen*

Compounds which contain a hyphen due to one or more constituents which are prefixations or combining forms (e.g. *auto-immune+ disease*) favour open spelling. → Tentatively confirmed.

Occasionally, prefixes are linked to a base by means of a hyphen, for example (but not necessarily) when there is a phoneme/grapheme clash at the boundaries, as in *pre-empt*. When such a hyphenated prefixation occurs as part of a compound, the presence of the hyphen in the compound shapes its visual appearance in such a way that this may influence what compound spelling variant is selected in order to express hierarchical structure: if we assume that the hyphen expresses a degree of proximity lying between the stronger one of solid spelling and the weaker one of open spelling, the choice of open spelling in a compound like *non-executive director* conveys immediately that the link between the prefix and its base is stronger than that between the constituents of the compound. By contrast, the solid alternative **non-executivedirector* would suggest the (incorrect) structure **non- + [executive+director]*, while hyphenation (*non-executive-director*) would suggest that the two hyphens have equal status and that all constituents are situated on the same structural level. We may therefore formulate the following expectation:

- E5: Compounds which contain a hyphen due to one or more constituents which are prefixations or combining forms (e.g. *auto-immune+disease*) favour open spelling.

In order to test this hypothesis, it was necessary to go beyond the OHS_600 set: the original LDOCE compound list included ten items containing a hyphenated prefixation, which were deleted from the final Master List (cf. 4.1). While their spelling may merely reflect a convention of the publisher, with the small number of items preventing statistical testing, a very obvious tendency was observable: all items with two freely occurring constituents used open spelling (*air vice-marshal*, *anti-virus software*, *auto-immune disease*, *extra-sensory perception*, *non-commissioned officer*, *non-executive director*, *semi-skimmed milk*) and all items with three freely occurring constituents framed the hyphenated constituent with a space (*anti-lock braking system*, *hand-eye co-ordination*, *post-traumatic stress disorder*).¹² Hypothesis E5 can thus be tentatively confirmed: compounds

¹² Possible hierarchical structures for triconstituent compounds are $[A+B]+C$ and $A+[B+C]$ – alternatively, $A+B+C$ when all constituents have equal status. Schmid (2011: 207–208) finds triconstituent noun compounds with a complex modifier, i.e. $[N + N] + N$, to be almost five times as frequent in his corpus as those with a complex head, i.e. $N + [N + N]$. He explains his observation that almost all complex heads are highly established compounds with solid spelling (e.g. *girlfriend*, *bedroom*) by the fact that the head as the grammatically, semantically and conceptually dominant constituent should be occupied by an established lexeme whose concept is stored in the mental lexicon (Schmid 2011: 208). At the same time, we can infer from the frequency differences between these two patterns that a structural misinterpretation of $[A+B]+C$

which contain hyphenated prefixations or combining forms as constituents favour open spelling. This preference for a spelling which reflects the strength of the links between the constituents was also observed by Bertram et al. (2011), who inserted illegitimate hyphens in Dutch and Finnish triconstituent compounds and found that the inclusion of a hyphen at the minor boundary in an otherwise solid compound generated “problems for detecting the hierarchical morphological structure and for integration of all constituents into a unified meaning” (Bertram et al. 2011: 537). However, the principle of hierarchical orthographic ordering is not binding, as can be observed in a list of recent additions to the *Oxford English Dictionary*, where *credit card-sized* uses open spelling in *credit card* by analogy to the triconstituent compounds *credit card number* and *credit card details* while still adding *sized* with a hyphen, as is commonly done elsewhere, e.g. in *hand-sized*.¹³

5.5.2 Morphological Structure

The morphological structure of a compound involves a variety of sub-phenomena such as the complexity of the bases and the ordering of the constituents.

5.5.2.1 Hypothesis E6 – Complex Constituent(s)

Compounds containing one or more complex constituents disfavour solid spelling. → Confirmed.

As we have seen (cf. 5.5.1.5), the constituents of compounds may themselves be complex, e.g. affixations, acronyms or compounds. In this case, the internal link between the parts of the constituents (e.g. a prefix and a base) should be stronger than the link between the constituents of the compound. In order to express this difference in hierarchical structure, spellers are likely to make use of more than one type of compound spelling at the constituent-internal and external joints. As a consequence, we may formulate the following expectation:

compounds such as *entry+level+job* is very unlikely even if all constituents are merely separated by blanks, whereas the A+[B+C] compounds such as *school+time+table* will clearly benefit from constituent joint marking of different strength in order to avoid a garden path reading.

¹³ Cf. <http://public.oed.com/the-oed-today/recent-updates-to-the-oed/previous-updates/march-2013-update/new-words-list-march-2013/> (last accessed 27 July 2017) and <http://public.oed.com/the-oed-today/recent-updates-to-the-oed/previous-updates/june-2013-update/new-words-list-june-2013/> (last accessed 27 July 2017).

Table 5.40 *Complex constituents [Complex_const] in OHS_600*

Code	Description	Frequency	Examples
	no complex constituent	408	<i>sun+rise</i>
p	prefixation	6	<i>no+nonsense</i>
s	lexical suffixation	141	<i>motion+sickness</i>
i	inflected form	31	<i>no+frills</i>
a	acronym	1	<i>zip+code</i>
x	more than one complex constituent	13	<i>safety+razor</i>

E6: Compounds containing one or more complex constituents disfavour solid spelling.

In order to test Hypothesis E6, prefixations, suffixations and acronyms occurring as compound constituents were coded in a separate column of the database, regardless of their position within the compound. Slight formal deviations (e.g. concerning the pronunciation and/or spelling of *business* and *pressure*, or irregular inflection) were admitted, but the constituents had to be synchronically recognisable and still in use. An additional code was introduced for the occurrence of more than one complex constituent in a compound, e.g. in twice-suffixed *creepy+crawly*.

In order to meet the requirements for statistical testing, the data were recoded by considering only a binary opposition between the presence and absence of complex constituents [Complex_const_r]. In view of the predominance of lexical suffixations among the complex constituents (with 141 counts as against 51 others), the spelling preferences of this type of complex compound may be overrepresented in the results.

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped complex constituents' [Complex_const_r] was highly significant ($p = 0.000$) for OHS_600, as the presence of complex constituents clearly disfavours solid spelling (2.6 per cent), while favouring open spelling (39.1 per cent) and particularly hyphenation (58.3 per cent). The preferences of the compounds without complex constituents are less marked, but there is still a very clear tendency towards solid spelling (47.8 per cent), accompanied by the avoidance of hyphenation (21.6 per cent). This seems to suggest that the distinction between complex and simple constituents is a very important and basic distinction with regard to compound spelling, and that the absence of complex constituents also has an effect on spelling. Hypothesis E6 can

Table 5.41 *Grouped complex constituents [Complex_const_r] and spelling in OHS_600*

			Complex constituent(s)		Total
			–	+	
OHS	o	Count	125	75	200
		Expected Count	136.0	64.0	200.0
		% within Complex_const_r	30.6%	39.1%	33.3%
	h	Count	88	112	200
		Expected Count	136.0	64.0	200.0
		% within Complex_const_r	21.6%	58.3%	33.3%
	s	Count	195	5	200
		Expected Count	136.0	64.0	200.0
		% within Complex_const_r	47.8%	2.6%	33.3%
	Total	Count	408	192	600
		Expected Count	408.0	192.0	600.0
		% within Complex_const_r	100.0%	100.0%	100.0%

therefore be confirmed: the presence of one or more complex constituents disfavours solid spelling. While this might be attributable either to the underlying structure or to the increase in length, Rakić (2009: 60) observes an effect of increased complexity that is independent from length for his noun+noun compounds.

5.5.2.2 *Hypothesis E7 – No Head-Final Morphological Structure*

Compounds with a morphological structure which is not head-final (e.g. *court+martial*) disfavour solid spelling. → Refuted.

The literature on word formation recognises different types of compound based on the presence or absence of a head (i.e. a hyperonymous concept explicitly contained within a constituent; cf. Huddleston and Pullum 2002: 1645). Thus the second constituent of the headed endocentric compound *tooth+ache* can roughly represent the compound’s meaning on its own (a toothache is a kind of ache), whereas the referent of an exocentric compound such as *guinea pig* is metaphorical (i.e. not a kind of pig), so that the latter can be considered grammatically headed but semantically unheaded (cf. Plag 2003: 145–148). Coordinate compounds like *singer-songwriter* also lack hierarchical structure, because they are connected by an AND relation and reversible in theory (though not necessarily in practice due to experiential salience; cf. Newman and Rice 2006: 236).

While Sepp (2006: 19) considers coordinate compounds unheaded, the present approach regards them as comprising two heads, since they designate entities that are two things at the same time (e.g. both a singer and a songwriter in *singer-songwriter*). Prototypical English compounds have a head-final structure (cf. e.g. Dressler 2005: 43), but under French influence it became possible to invert this usual sequence, e.g. in *court-martial* and *Lords Temporal* (Faiß 1992: 74). Certain patterns emerge from the spelling of the example words in different grammars: thus compounds with *in+law* (e.g. *mother-in-law*) or *elect* (e.g. *president-elect*) are often hyphenated (cf. e.g. Strumpf and Douglas 1988: 50, 56), which might be attributable to the initial position of their head. Furthermore, the majority of reference works (e.g. Quirk et al. 1985: 1569) seem to agree on the hyphenation of phrase compounds like the adjective *do-it-yourself*, which are not hierarchically structured and thus unheaded. These observations culminate in the following hypothesis:

- E7: Compounds with a morphological structure which is not head-final disfavour solid spelling.

In order to test this hypothesis, word structure was coded manually for the OHS_600 compounds: head-final structures (as the unmarked case) with the code *f*; head-initial structures with the code *i* and non-headed (i.e. exocentric) structures with the code *n*. Compounds whose constituents can both stand for the meaning of the whole compound (e.g. *hemline*, which is both a hem and a line) were considered two-headed and coded *t* accordingly. Where the part of speech of the head and that of the compound differ (e.g. in the case of the noun *bust-up*, which is strongly related to the meaning of its initial verbal constituent *to bust*), the compound was classified as non-headed and additionally marked with the code *_pos* if it was not possible to integrate the potential head into a paraphrase corresponding to the specific part of speech (e.g. *‘a bust-up is a kind of to bust’).

The distinction between non-headed and head-final compounds resulted in some borderline cases for compounds with non-literal meaning components: some compounds (e.g. *fire+wall*) have both a literal meaning (‘a special wall that prevents fires from spreading to other parts of a building’; cf. LDOCE for all meaning descriptions) and a more metaphorical meaning (‘a system that protects a computer network from being used or looked at by people who do not have permission to do so’). While *wall* is the obvious morphological and semantic head for the ‘building’ meaning of *fire+wall*, the ‘computer’ meaning requires a metaphorical extension of

Table 5.42 *Morphological structure [Morphol_struct] in OHS_600*

Code	Description	Frequency	Frequency by part of speech	Examples
f	head-final	462	356 n 98 adj 6 v 2 adv	<i>hand+book</i> 'a kind of book'
i	head-initial	20	15 adj 5 n	<i>broken+down</i> 'broken'
n	non-headed	69	56 v 8 adj 2 adv 3 n	<i>bell+bottoms</i> 'trousers'
n_pos	non-headed due to violation of part-of- speech constraint	36	29 adj 4 v 2 n 1 adv	<i>all+star</i> (adj)
t	two-headed	13	7 adj 6 n	<i>hem+line</i>

wall (literally 'an upright flat structure made of stone or brick, which divides one area from another or surrounds an area'), which is provided by the function of keeping entities apart. Since the first constituent in such compounds still modifies the second constituent (which can be considered a hyperonym in the widest sense), they were classified as morphologically headed. By contrast, in bahuvrihi compounds such as *paper+back*, the category of the potential head 'back' is not extended metaphorically, and the whole compound stands for another, unexpressed entity. Since the meaning of *paper+back* would rather be paraphrased as '[paper + back] book', *book* is regarded as the unexpressed, underlying head and bahuvrihi compounds are therefore classified as unheaded.

Table 5.42 gives an overview of the morphological structures which can be found in the OHS_600 compounds. While the majority of the head-final OHS_600 compounds in Table 5.42 are nouns, the concept of headedness can also be applied to compounds with other parts of speech (some of which belong at the periphery of the compound concept):

Table 5.43 *Grouped morphological structure [Morphol_struct_r] in OHS_600*

			Morphological structure		Total
			head-final	non-head-final	
OHS	o	Count	192	8	200
		Expected Count	154.0	46.0	200.0
		% within Morphol_struct_r	41.6%	5.8%	33.3%
	h	Count	108	92	200
		Expected Count	154.0	46.0	200.0
		% within Morphol_struct_r	23.4%	66.7%	33.3%
	s	Count	162	38	200
		Expected Count	154.0	46.0	200.0
		% within Morphol_struct_r	35.1%	27.5%	33.3%
	Total	Count	462	138	600
		Expected Count	462.0	138.0	600.0
		% within Morphol_struct_r	100.0%	100.0%	100.0%

- **adjectives:** the first constituent of *all+important* further specifies the quality of the second one; *earth+shattering* confessions are extremely shattering
- **verbs:** someone who *chain+smokes* has a particular way of smoking as a habit, and if you *double+check* something, you check it twice
- **adverbs:** *well+nigh* can be replaced by *nigh*; *stage left* is a subtype of 'left'
- **prepositions:** the LDOCE examples for *onto* can either be replaced by *on* (*pour the syrup on(to) the egg mixture*) or by *to* (*a gate leading (on)to a broad track*); *upon* can be replaced by *on* in many contexts (e.g. *to be dependent (up)on sb*) but e.g. not in **once on a time*
- **interjections:** *oops* can be used in place of *oops-a-daisy*; *heave* can be used for *heave-ho*
- **pronouns:** *such* on its own can roughly transmit the meaning of *such+and+such*
- **conjunctions:** *as* can replace *in+as+much+as* in the LDOCE example *Ann is guilty, inasmuch as she knew what the others were planning.*

In order to test Hypothesis E7, Pearson's chi-square test was carried out for the dependent variable 'compound spelling' [OHS] and the independent variable 'presence/absence of a head-final morphological structure' [Morphol_struct_r] for the OHS_600 sample.

The result was highly significant ($p = 0.000$) due to the non-head-final constructions' clear avoidance of open spelling (5.8 per cent) and their clear preference for hyphenation (66.7 per cent). By contrast, the

proportion of 27.5 per cent solid spellings is so close to the 33.3 per cent which one would have expected by chance that Hypothesis E7 must be refuted: compounds with a morphological structure which is not head-final do therefore not disfavour solid spelling. The same picture emerges if we consider the subtypes of morphological structure (cf. Table 5.42), which meet the requirements for statistically valid testing. Pearson's chi-square test for OHS_600 with the dependent variable 'compound spelling' [OHS] and the independent variable 'morphological structure' [Morphol_struct] yielded highly significant results ($p = 0.000$). All non-head-final categories agree in their clear preference for hyphenation and their clear avoidance of open spelling (cf. Table 5.44). The most extreme results are obtained for the head-initial compounds (i), with 0 per cent open spellings and 80 per cent hyphenations. By analogy to Sepp's (2006: 113) explanation for hyphenation in compounds containing *-er* or *-ing* (cf. 5.5.1.4), one might argue that the recognition and analysis of non-canonical ordering of the constituents in non-head-final compounds is aided by hyphenation. However, since this still fails to explain the even more extreme avoidance of open spelling, the postulation of a hyphenation principle for marked compounds might make even more sense (cf. 7.1).

5.5.3 Summary

Section 5.5 investigates morphology-related variables which were expected to exert some influence on the spelling of English biconstituent compounds.

The following variables were coded in the database:

- Occurrence of prefix following the constituent joint (*no+nonsense*)
- Occurrence of the suffixes *-ing*, *-ed* or *-er* in compound-final position (*race+meeting*)
- Occurrence of lexical suffix preceding the constituent joint (*poetic+licence*)
- Occurrence of inflection preceding the constituent joint (*systems+analyst*)
- Complex constituents (e.g. prefixations, acronyms)
- Morphological structure (e.g. head-initial, head-final, non-headed).

Statistical testing revealed a strong effect of several independent variables on the dependent variable 'compound spelling' [OHS] for the OHS_600 compounds. The following hypotheses were confirmed:

Table 5.44 Morphological structure [Morphol_struct] and spelling in OHS_600

		Morphological structure						Total
		f	i	n	n_pos	t		
OHS	o	Count	192	0	7	0	1	200
		Expected Count	154.0	6.7	23.0	12.0	4.3	200.0
		% within Morphol_struct	41.6%	0.0%	10.1%	0.0%	7.7%	33.3%
h		Count	108	16	41	28	7	200
		Expected Count	154.0	6.7	23.0	12.0	4.3	200.0
		% within Morphol_struct	23.4%	80.0%	59.4%	77.8%	53.8%	33.3%
s		Count	162	4	21	8	5	200
		Expected Count	154.0	6.7	23.0	12.0	4.3	200.0
		% within Morphol_struct	35.1%	20.0%	30.4%	22.2%	38.5%	33.3%
Total		Count	462	20	69	36	13	600
		Expected Count	462.0	20.0	69.0	36.0	13.0	600.0
		% within Morphol_struct	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

- Non-compound-final lexical suffixes disfavour solid spelling. [E1]
- The presence of the compound-final suffixes *-ing*, *-ed* or *-er* favours hyphenation. [E4]
- One or more complex constituents disfavour solid spelling. [E6]

By contrast, the result for the following hypothesis contradicts the expectations:

- A morphological structure which is not head-final does not disfavour solid spelling. [E7]

Although the following results are not backed by statistical significance, a number of tendencies can be observed:

- Non-compound-final grammatical suffixes disfavour solid spelling. [E2]
- Non-compound-initial prefixes disfavour solid spelling. [E3]
- Compounds containing a hyphen due to one or more constituents which are prefixations or combining forms favour open spelling. [E5]

Some of the results for the more general hypotheses, which merely predict the avoidance of one particular spelling, permit more precise conclusions than expected for a number of variables:

- Non-compound-final lexical suffixes favour open spelling. [E1]
- Complex constituents favour open spelling and particularly hyphenation. [E6]

In addition to the results for the hypotheses under consideration, several other findings emerged from the detailed analysis of the material:

- Compounds which do not contain non-final lexical suffixes avoid open spelling – but only very slightly. [E1]
- The presence of the compound-final suffixes *-ing*, *-ed* or *-er* disfavour solid spelling. [E4]
- A morphological structure which is not head-final favours hyphenation and disfavour open spelling. [E7]

5.6 Grammar

While grammar is prototypically concerned with linguistic units larger than the word, some aspects of grammar need to be considered as potential determinants of variant selection in English compound spelling. The

following sections therefore discuss the potential influence of part of speech, syntactic context and compound-internal clause structure.

5.6.1 Part of Speech

Part of speech may influence the spelling of English compounds on two levels: with regard either to the part of speech of the whole compound or to the part of speech of its constituents (including their combination).

5.6.1.1 Hypothesis F₁ – Part of Speech of the Compound

The part of speech of the compound influences its spelling. → Confirmed.

The part of speech of the compound traditionally plays an important role in English compound spelling as a selection criterion for the material to be analysed in all previous linguistic studies (which focus on noun compounds; cf. 1.1.1). Furthermore, style guides commonly structure their advice first by the part of speech of the compound to be spelled and then by the part of speech of its constituents (e.g. Peters 2004; Merriam-Webster 2001). There is also an observable tendency for compounds that are identical except for their part of speech to be spelled differently: *spot+check* is spelled open as a noun (*The spot check was a surprise*) but hyphenated as a verb (*If you can't proofread the document, at least spot-check it*; cf. Wilbers 1997), the hyphenated adjective *face-to-face* contrasts with the open adverb *face to face*, and *main+line* is spelled open as a noun but solid as an adjective (Macmillan 2007). Since all of this suggests that part of speech plays some role in the determination of English compound spelling, this can be formulated as Hypothesis F₁:

F₁: The part of speech of the compound influences its spelling.

No additional coding was necessary for the testing of Hypothesis F₁, since the part of speech of the compounds was already contained in the LDOCE-based Master List (cf. 4.1).

OHS_600 contains 426 nouns (71.0 per cent), 157 adjectives (26.2 per cent), 12 verbs (2.0 per cent) and 5 adverbs (0.8 per cent), which is similar to the proportion of parts of speech in the Master List (cf. Table 5.45) comprising all compounds.¹⁴ The expected influence of part of speech on

¹⁴ While the approach followed here discusses the items listed in Table 5.45 as grammatical compounds (cf. Chapter 2 and Section 5.6.1), an alternative analysis in terms of valency grammar (e.g. Herbst

Table 5.45 *Part of speech of the whole compound [PoS_comp] in the Master List*

Part of speech	Lexical/ grammatical	Frequency	%	Grammatical compounds (complete list)
n	lex	8,536	85.01	
adj	lex	1,298	12.93	
v	lex	136	1.35	
adv	lex	51	0.51	
prep	gr	10	0.10	<i>according+to</i> <i>apart+from</i> <i>due+to</i> <i>in+to</i> <i>next+to</i> <i>on+to</i> <i>owing+to</i> <i>relating+to</i> <i>through+out</i> <i>up+on</i>
interj	gr	5	0.05	<i>aw+shucks</i> <i>full+stop</i> <i>heave+ho</i> <i>oops+a+daisy</i> <i>thank+you</i>
pron	gr	4	0.04	<i>each+other</i> <i>no+one</i> <i>one+another</i> <i>such+and+such</i>
conjunction	gr	1	0.01	<i>in+as+much+as</i>
TOTAL		10,041	100	

the spelling of English compounds was confirmed by Pearson’s chi-square test for the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘grouped part of speech of the compound’ [PoS_comp_r] (in which verbs and adverbs were treated jointly due to their small numbers) for OHS_600 ($p = 0.000$). The highly significant results are due to the absence of open spellings in adjectives, adverbs and verbs, but also to the distributional pattern of the noun compounds: these are either spelled open (46.9 per cent) or solid (42.5 per cent) but rarely hyphenated (10.6 per cent). Hypothesis F1 can thus be confirmed: the part of speech of the compound influences its spelling.

and Schüller 2008) may consider some of these items, particularly the prepositions and the conjunction, as valency patterns.

Table 5.46 *Grouped part of speech of the whole compound [PoS_comp_r] and spelling in OHS_600*

			Part of speech			Total
			adj	n	v / adv	
OHS	o	Count	0	200	0	200
		Expected Count	52.3	142.0	5.7	200.0
		% within PoS_comp_r	0.0%	46.9%	0.0%	33.3%
	h	Count	144	45	11	200
		Expected Count	52.3	142.0	5.7	200.0
		% within PoS_comp_r	91.7%	10.6%	64.7%	33.3%
	s	Count	13	181	6	200
		Expected Count	52.3	142.0	5.7	200.0
		% within PoS_comp_r	8.3%	42.5%	35.3%	33.3%
Total	Count		157	426	17	600
	Expected Count		157.0	426.0	17.0	600.0
	% within PoS_comp_r		100.0%	100.0%	100.0%	100.0%

5.6.1.2 Hypothesis F2 – Adjective Compounds

Adjective compounds favour hyphenation. → Confirmed.

Adjectives come after noun compounds regarding research on their spelling (cf. 1.1.1). Since style guides commonly recommend hyphenation as the default spelling variant for this group of compounds (e.g. Peters 2004: 259; Strumpf and Douglas 1988: 49), we may formulate the following expectation:

F2: Adjective compounds (e.g. *ill+advised*) favour hyphenation.

The discussion of Hypothesis F2 is based on the significant results ($p = 0.000$) for Hypothesis F1, which are summarised in Table 5.46 and partly due to the behaviour of the adjective compounds: with no open spellings, 8.3 per cent solid spellings and 91.7 per cent hyphenations, Hypothesis F2 can be confirmed: adjective compounds favour hyphenation indeed. This result is in line with Mondorf (2009: 378), who finds hyphenation in about 85 per cent of adjectival compounds. In this type of compound, we often observe combinations of constituents that do not comply with usual syntactic structures, e.g. the order *n+adj* (*ice+cold*) compared to the usual order *adj+n* in the noun phrase (*cold ice*; cf. also Sepp 2006: 113), or the missing determiner in the adjectives *in+store* and *off+centre* by comparison to the noun phrases *in the store* and *off the centre*. That might be the reason

why hyphenation as the most marked spelling is so commonly used in adjective compounds.

5.6.1.3 Hypothesis F3 – Adjectives with Compound-Initial -ly Adverb

Adjective compounds with a compound-initial adverb ending in *-ly* (e.g. *neatly+dressed*) favour open spelling. → Tentatively confirmed.

However, there are some exceptions to the default hyphenation of adjectives noted in Section 5.6.1.2: solid spelling is said to be found in very well-established adjective compounds consisting of a simple adverb followed by a participial form (e.g. *everlasting*, *widespread*; cf. Peters 2004: 259), and one extremely common orthographic rule in the literature is that adjectival compounds with a first constituent ending in adverbial *-ly* should be spelled open, e.g. *politically correct* (*opinions*) (cf. Merriam-Webster 2001: 105; Quirk et al. 1985: 1613–1614; *GPO Style Manual* 2008: 78). Goldstein (2004: 332) explains that no hyphen is required in such compounds because readers can expect the *-ly* adverb to modify the following word. However, some exceptions can be found in the literature (e.g. commonly hyphenated *newly-wed*; cf. Ritter 2005a: 54), and the principle was therefore subjected to empirical testing of Hypothesis F3:

F3: Adjective compounds with a compound-initial adverb ending in *-ly* (e.g. *neatly+dressed*) favour open spelling.

Since OHS_600 does not contain any compounds corresponding to the required pattern, additional compound lists were drawn upon: OHS_extra contains three fitting compounds (*genetically modified*, *partially sighted* and *politically correct*), all of which are always spelled open, as expected. To this can be added the only compound of this type from Master_5+, *fully+fledged*, which occurs three times with open spelling but also three times with hyphenation. Among the compounds with up to four occurrences in the dictionaries, we also find a clear predominance of open spelling for this group of compounds (cf. Table 5.47).

As a consequence, Hypothesis F3 can be tentatively confirmed, even if statistical backing is lacking: adjective compounds with a compound-initial adverb ending in *-ly* favour open spelling, as advocated by the style guides and grammars. This clearly distinguishes them from adjectives in general (with 0 per cent open spellings; cf. 5.6.1.2). Furthermore, adjectives with an initial *-ly* adverb have a stronger tendency than adjectives in general to avoid solid spelling (with a proportion of 0 per cent vs. 8.3 per cent for all adjectives; cf. 5.6.1.2). This special behaviour can be explained by the

Table 5.47 *Adjective compounds with an initial -ly adverb in Master_1–4*

Compound	Open	Hyphenated	Solid
<i>badly+off</i>	3	1	0
<i>downwardly+compatible</i>	1	0	0
<i>downwardly+mobile</i>	2	0	0
<i>environmentally+friendly</i>	4	0	0
<i>fully+dressed</i>	1	0	0
<i>fully+grown</i>	0	2	0
<i>mentally+handicapped</i>	4	0	0
<i>mentally+ill</i>	1	0	0
<i>physically+challenged</i>	2	0	0
<i>politically+incorrect</i>	3	0	0
<i>poorly+off</i>	2	0	0
<i>softly+spoken</i>	1	2	0
<i>tightly+knit</i>	0	2	0
<i>upwardly+mobile</i>	4	0	0

presence of the suffix right before the constituent joint, which disfavours solid spelling (cf. 5.5.1.1).

5.6.1.4 Hypothesis F4 – Verb Compounds

Verb compounds disfavour open spelling. → Tentatively confirmed.

No reference work seems to consider the possibility of open spelling in compound verbs (cf. e.g. Partridge 1953: 146 or the *GPO Style Manual* 2008: 83). Peters (2004: 259) implicitly excludes open spelling when stating that verbal compounds are “either hyphenated or set solid, depending on their components”. All of the foregoing information enters Hypothesis F4:

F4: Verb compounds (e.g. *force+feed*) disfavour open spelling.

While it was not possible to obtain statistically valid results, since there are only twelve verb compounds in OHS_600, the occurrence of zero open, nine hyphenated and three solid spellings among these lends support to Hypothesis F4. The additional analysis of the forty-eight verb compounds in OHS_extra confirms this result: with eighteen hyphenations and thirty solid spellings contrasting with the complete avoidance of open spelling, Hypothesis F4 – that verb compounds disfavour open spelling – can be tentatively accepted. A possible explanation for this finding is the status of

verbs as the central syntactic elements, as postulated by valency theory (cf. e.g. Herbst et al. 2004; Herbst and Schüller 2008). Since the interpretation of the whole sentence depends on the verb, the spelling of verbs as uninterrupted chains of characters supports their identification as a single unit and thereby represents an advantage for processing.¹⁵

5.6.1.5 Hypothesis F₅ – Adverb Compounds

Adverb compounds disfavour open spelling. → Tentatively confirmed.

According to Peters (2004: 259), compound adverbs (e.g. *barefoot, down-stairs*) “are usually set solid”, and Merriam-Webster (2001: 109) states that “[a]dverb compounds consisting of preposition + noun are almost always written solid”. Partridge (1953: 147), by contrast, claims that the majority of compound adverbs are hyphenated (e.g. *They collided head-on*). Jointly, this permits the formulation of the following hypothesis for biconstituent compounds:

F₅: Adverb compounds disfavour open spelling.

OHS_600 contains only five adverb compounds (which precludes statistical testing), but it is striking that none of these adverbs is spelled open, whereas three use solid spelling and two are hyphenated. The larger OHS_extra list contains a majority of eight solid adverbs and one hyphenation, but also two adverbs with open spelling (*next door; upside down*). Of the four adverb compounds in Master_5+, two types (*no+place; double+quick*) have a total of five open spellings in the dictionaries. With four tokens, open spelling is even favoured for *no+place*, but altogether, open spelling is less frequent with five tokens as against nine hyphenated and six solid spellings. The prototypical adverbs (in OHS_600 and OHS_extra) thus seem to disfavour open spelling – which would tentatively support Hypothesis F₅ – but if we extend the category to the cases with spelling variation, the picture becomes less clear.

5.6.1.6 Hypothesis F₆ – Grammatical Compounds

Grammatical compounds disfavour open spelling. → Tentatively refuted.

¹⁵ Since phrasal and prepositional verbs always use open spelling (cf. 2.3), this might point to an important difference regarding processing between these two types of complex verb compared to verbal compounds.

Table 5.48 *Spelling of the adverb compounds in Master_5+*

Compound	O	H	S
<i>double+quick</i>	1	4	0
<i>no+place</i>	4	0	1
<i>post+haste</i>	0	2	3
<i>side+saddle</i>	0	3	2

Grammatical parts of speech are rarely the product of compounding – which might explain why they are barely mentioned in the literature on compound spelling and commonly ignored in the sections on compounding in grammars such as Huddleston and Pullum (2002) or Quirk et al. (1985). Nonetheless, prepositions, pronouns, conjunctions, numerals, determiners and interjections may also exhibit characteristic spellings: thus Ritter (2005a: 56) advocates the use of hyphens “in spelled-out numbers from 21 to 99”, e.g. *twenty-nine*. According to Partridge (1953: 143), most compound pronouns (e.g. *anybody*) are written solid, and Morton Ball (1939: 76) advocates solid spelling for “all compound pronouns, prepositions, and conjunctions”. Taking all of this into account, one may formulate the following hypothesis:

F6: Grammatical compounds disfavour open spelling.

In order to test Hypothesis F6, the binary distinction between compounds with lexical and grammatical parts of speech (cf. Quirk et al. 1985: 72) was coded semi-automatically based on the part of speech of the whole compound.

Since the OHS_600 list contains no grammatical compounds, OHS_extra was drawn upon, which contains four prepositions and two pronouns. Of these, the prepositions *into*, *throughout* and *upon* are spelled solid, but *according to* and the pronouns *each other* and *one another* all have exclusively open spelling. Master_5+ contains one pronoun (*no+one* with six open spellings and one hyphenation) and one preposition (*on+to* with one open and six solid spellings). In spite of the small number of grammatical compounds in the data, which prevents statistically valid results, Hypothesis F6 can be tentatively refuted: grammatical compounds do not disfavour open spelling.

5.6.1.7 *Hypothesis F7 – Part of Speech of the Constituents*

The part of speech of the constituents influences the spelling of the compound. → Confirmed.

Besides the level of the compound as a whole, part of speech may also concern the constituents. Carey (1957: 25–26) argues that compounds with premodifying adverbs (e.g. *neatly+dressed*) need no hyphen, since adverbs are “tacked on” by nature (Carey 1957: 26). The present study recognises English compounds with a large number of possible part-of-speech combinations (cf. Table 2.4) and is not limited to nominal noun+noun compounds like most of its predecessors. The expectation underlying this approach can be formulated in the following way:

F7: The part of speech of the constituents influences the spelling of the compound.

In order to test this hypothesis, the part of speech of each constituent was coded for the OHS_600 compounds in separate columns of the database. However, the part of speech of a compound’s constituents cannot always be determined unequivocally, e.g. in the example *rattlesnake*, in which *rattle* might be classified either as a verb (*the snake rattles*) or as a noun (*the snake has a rattle*; cf. Bauer 1983: 202). Part of speech is traditionally determined by using the position of a lexeme within the context of a phrase or sentence, as well as the sparse inflectional morphology English has to offer. Since the part of speech of compound heads frequently corresponds to that of the whole compound, part-of-speech assignment is more problematic for non-heads. In general, part of speech within a compound can only be assigned based on a clausal paraphrase (cf. also 5.6.3).

In the codification, an attempt was made to assign the ‘default’ part of speech to the constituents. Thus both constituents of the noun *brick wall* are classified as nouns – in spite of the fact that both *brick* and *wall* may be verbs in other contexts (cf. LDOCE 2009) and regardless of the fact that some grammars, such as Quirk et al. (1985: 1562), consider *brick* an adjective in this context because it can also occur predicatively in *This wall is brick*. Constituents with doubtful part of speech were classified according to the part of speech used in the description of the compound’s meaning in the reference works LDOCE and OED. For instance, *drum* in *drum machine* was coded as a noun (rather than a verb) based on the LDOCE paraphrase ‘a piece of electronic equipment that makes patterns of sounds like drum music’.

Pushing the approaches in Herbst and Schüller (2008) or Langacker (2008) even further, the present study recognises the following parts of speech for the constituents of compounds:

noun (n), e.g. *air* in *air+plane*
 verb (v), e.g. *bail* in *bail+out*
 adjective (adj), e.g. *bad* in *bad+guy*
 adverb (adv), e.g. *badly* in *badly+off*.

All other parts of speech – e.g. traditional prepositions, conjunctions, numerals and pronouns – were categorised as *g* for *grammatical*, e.g.

at in *at+ risk* (adj)
up in *well+brought+up* (adj)
if in *what+if* (n)
what in *what+if* (n)
the in *jack+in+the+box* (n)
you in *thank+you* (n)
sixth in *sixth sense* (n).

The reason for this treatment of grammatical words as a single class is that strictly speaking, there is no grammatical context which could be used for the determination of part of speech within compounds – except in a few special cases such as phrase compounds or possibly genitive compounds. Since grammatical words such as *before*, *up* and *until* frequently permit several possible uses in syntactic structures, e.g. as adverbs, prepositions or conjunctions (cf. Herbst and Schüller 2008: 67), the lack of context makes it impossible to assign a precise part of speech to them based on their distribution within a compound. Furthermore, they lend themselves less well than the lexical words to a prototype-based part-of-speech classification which postulates that nouns typically refer to persons or things, adjectives to qualities, verbs to activities and adverbs to the way in which actions are carried out (cf. Herbst and Schüller 2008: 34 for a critical discussion of this traditional approach). If one further considers that most language users seems to identify nouns, adjectives and verbs relatively well while feeling far less certain about the classification of the usually short and highly frequent grammatical words (Herbst and Schüller 2008: 73), the subsumption of all these grammatical words into one category seems to make sense indeed for the purpose of the present study.

Traditional adverbs are often regarded as a problematic word class, because they constitute a very heterogeneous category (Quirk et al. 1985: 438) on a gradient between lexical and grammatical words, containing

items as diverse as *very* and *here*, which are directly opposed in their syntactic distribution, e.g. regarding the premodification of adjectives or the occurrence at the end of a clause as a complement of the verb *be* (Herbst and Schüller 2008: 59–61). For that reason, it was decided to restrict the category of adverbs in the present study, which has a lexical focus, to morphologically complex adverbs, including both derived (*badly*) and compound adverbs (*hereby*). This avoids the part-of-speech ambiguity which seems more common in the morphologically simple adverbs (cf. e.g. Quirk et al. 1985: 438).

Since the compounds in OHS_600 consist of two constituents, these were considered separately in order to test Hypothesis F7. First, Pearson's chi-square test was carried out for the dependent variable 'compound spelling' [OHS] and the independent variable 'part of speech of the first constituent' [PoS_1] for OHS_600. The results are highly significant ($p = 0.000$) due to the following correlations between part of speech of the first constituent and spelling:

- A compound-initial adjective favours open spelling (50.0 per cent) and disfavors solid spelling (20.9 per cent). This is striking in view of the fact that open spelling makes nominal adj+n compounds indistinguishable from phrases with the same pattern. The orthographic distinction between compound and phrase exemplified in the classic textbook example *blackbird* (= solid compound) vs. *black bird* (= open phrase) should consequently not be overestimated in its importance.
- A compound-initial adverb favours hyphenation (94.9 per cent).
- A compound-initial grammatical word favours hyphenation (87.1 per cent).
- A compound-initial noun disfavors hyphenation (15.9 per cent). The proportions of solid and open spelling are very close for this part of speech (43.5 per cent and 40.6 per cent, respectively).
- A compound-initial verb favours hyphenation (56.6 per cent).

In order to meet the requirements for statistical testing, the part of speech of the second constituent had to be recoded by treating verbs and adverbs jointly [PoS_2_1]. The result of Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped part of speech of the second constituent' [PoS_2] was highly significant ($p = 0.000$) due to the following correlations between part of speech of the second constituent and spelling:

Table 5.49 *Part of speech of the first constituent [PoS_1] and spelling in OHS_600*

		Part of speech of the first constituent					Total
		adj	adv	g	n	v	
OHS	o	Count	0	2	141	2	200
		Expected Count	19.7	10.3	115.7	17.7	200.0
		% within PoS_1	0.0%	6.5%	40.6%	3.8%	33.3%
	h	Count	56	27	55	30	200
		Expected Count	19.7	10.3	115.7	17.7	200.0
		% within PoS_1	94.9%	87.1%	15.9%	56.6%	33.3%
Total	s	Count	3	2	151	21	200
		Expected Count	19.7	10.3	115.7	17.7	200.0
		% within PoS_1	5.1%	6.5%	43.5%	39.6%	33.3%
		Count	59	31	347	53	600
		Expected Count	59.0	31.0	347.0	53.0	600.0
		% within PoS_1	100.0%	100.0%	100.0%	100.0%	100.0%

Table 5.50 *Grouped part of speech of the second constituent [PoS_2_r] and spelling in OHS_600*

			Part of speech of the second constituent				Total
			adj	g	N	v / adv	
OHS	o	Count	0	2	197	1	200
		Expected Count	36.0	13.7	144.0	6.3	200.0
		% within PoS_2_r	0.0%	4.9%	45.6%	5.3%	33.3%
	h	Count	99	35	49	17	200
		Expected Count	36.0	13.7	144.0	6.3	200.0
		% within PoS_2_r	91.7%	85.4%	11.3%	89.5%	33.3%
	s	Count	9	4	186	1	200
		Expected Count	36.0	13.7	144.0	6.3	200.0
		% within PoS_2_r	8.3%	9.8%	43.1%	5.3%	33.3%
Total	Count		108	41	432	19	600
	Expected Count		108.0	41.0	432.0	19.0	600.0
	% within PoS_2_r		100.0%	100.0%	100.0%	100.0%	100.0%

- A compound-final adjective favours hyphenation (91.7 per cent).
- A compound-final grammatical word favours hyphenation (85.4 per cent).
- A compound-final noun disfavors hyphenation (11.3 per cent). The proportions of open spelling (45.6 per cent) and solid spelling (43.1 per cent) are very close for this part of speech.
- A compound-final verb or adverb favours hyphenation (89.5 per cent). This result does not change if the two parts of speech are considered individually, and it was therefore included in the list of significant variables in the Appendix (cf. Table A.9).

Since practically all first and second constituents show a very clear tendency in favour of one of the spelling variants, Hypothesis F7 can be confirmed: the part of speech of the constituents influences the spelling of the compound. Merely the results for the nouns are slightly inconclusive in both positions: the only clear tendency is the avoidance of hyphenation, whereas open and solid spellings are practically equally likely. The dispreference for hyphenation in this type of compound, which is the most prototypical one judging from the number of studies devoted to it (cf. 1.1.1), may have contributed to the view that hyphenation is unusual and to the prescriptive notion that hyphenation should be avoided. However, the view of hyphenation as unusual is incorrect considering the large number

Table 5.51 *Part-of-speech combinations for the first and second constituents in OHS_600 sorted by spelling variant*

	PoS combination	Frequency	Example
O	adj+n	55	<i>sour cream</i>
	g+n	2	<i>first class</i>
	n+adv	1	<i>centre forward</i>
	n+g	1	<i>number one</i>
	n+n	139	<i>time zone</i>
	v+g	1	<i>sod all</i>
	v+n	1	<i>tumble drier</i>
H	adj+adj	6	<i>red-faced</i>
	adj+g	4	<i>grown-up</i>
	adj+n	22	<i>low-key</i>
	adv+adj	51	<i>new-found</i>
	adv+adv	1	<i>well-nigh</i>
	adv+g	2	<i>close-up</i>
	adv+v	2	<i>double-check</i>
	g+adj	5	<i>all-important</i>
	g+g	3	<i>in-between</i>
	g+n	17	<i>no-frills</i>
	g+v	2	<i>second-guess</i>
	n+adj	36	<i>camera-shy</i>
	n+g	7	<i>hands-off</i>
	n+n	8	<i>bell-bottoms</i>
	n+v	4	<i>spoon-feed</i>
	v+adj	1	<i>drip-dry</i>
	v+adv	3	<i>go-ahead</i>
	v+g	19	<i>know-all</i>
	v+n	2	<i>glow-worm</i>
	v+v	5	<i>stop-go</i>
S	adj+n	23	<i>greyhound</i>
	adv+adj	2	<i>southbound</i>
	adv+g	1	<i>nearby</i>
	g+adj	1	<i>inbred</i>
	g+n	1	<i>indeed</i>
	n+adj	6	<i>moonlit</i>
	n+n	144	<i>workday</i>
	n+v	1	<i>hoodwink</i>
	v+g	3	<i>lookout</i>
	v+n	18	<i>swimsuit</i>

of compound types based on part-of-speech combination which are exclusively hyphenated in OHS_600 (cf. Table 5.51).

Since the constituents of a compound might favour different spelling variants based on their part of speech, it is interesting to consider what

Table 5.52 *Predicted spelling preferences of the OHS_600 compounds based on the part of speech of the first constituent [PoS_1] and the part of speech of the second constituent [PoS_2]*

second const. first const.	n	adj	v	adv	G
N	-h/-h	-h/h	-h/h	-h/h (o)	-h/h
Adj	o/-h	o/h	o/h	o/h	o/h
V	h/-h (s)	h/h	h/h	h/h	h/h
Adv	h/-h	h/h	h/h	h/h	h/h
G	h/-h	h/h	h/h	h/h	h/h

predictions emerge from their combination. The first value in each cell of Table 5.52 refers to the spelling preferred by the first constituent; the value following the slash represents the most common spelling for the second constituent. Codes preceded by a minus sign signify avoidance and shaded cells indicate contradictory tendencies of the two constituents. The spelling variant in bold print is the one preferred by the compounds with the corresponding part-of-speech combination in the OHS_600 data, the variant in brackets being a specification of ‘not hyphenated’. If no variant is highlighted, no OHS_600 compound comprised that specific part-of-speech combination, or the data were inconclusive.

Table 5.52 shows that fourteen of the twenty-five combinations do not contradict each other (e.g. both an initial verb and a final adjective favour hyphenation). Combinations of noun and noun definitely disfavour hyphenation, but it remains unclear whether the outcome is rather spelled open or solid. In the majority of contradictory combinations, the spelling of the second constituent was preferred. This might be explained by a tendency for the part of speech of the second constituent to be identical with that of the whole compound (which is the case of 500, i.e. 83 per cent, of the 600 OHS_600 compounds).

5.6.1.8 *Hypothesis F8 – Combination of Lexical and Grammatical Constituents*

Compounds containing both lexical and grammatical constituents
disfavour solid spelling. → Confirmed.

As part of the more general hypothesis that speakers are reluctant to concatenate very dissimilar constituents (cf. 5.12.2), it was assumed that part of speech – and in particular the binary distinction between lexical and

Table 5.53 *Part of speech of the first constituent [PoS_1] and the second constituent [PoS_2] of the OHS_600 compounds with grouping as lexical/grammatical*

Part of speech	Lexical/grammatical	Constituent 1		Constituent 2	
		Frequency	%	Frequency	%
adj	lex	110	18.3	108	18.0
adv	lex	59	9.8	5	0.8
g	gr	31	5.2	41	6.8
n	lex	347	57.8	432	72.0
v	lex	53	8.8	14	2.3

grammatical constituents (cf. 5.6.1.1 and 5.6.1.7) – may have an effect on spelling:

F8: Compounds containing both lexical and grammatical constituents disfavour solid spelling.

In order to test this hypothesis, the parts of speech of each constituent of the OHS_600 compounds (which had been coded for Hypothesis F7) were recoded as either lexical (n+v+adj+adv) or grammatical (cf. Table 5.53). There are only thirty-one compounds with a grammatical constituent in initial position (5.2 per cent), forty-one with a grammatical constituent in final position (6.8 per cent) and three lexical compounds composed of two grammatical constituents (the adverb *fifty+fifty*, the adjective *in+between* and the noun *once+over*).

Pearson's chi-square test was carried out for the dependent variable 'compound spelling' [OHS] and the independent variable 'combination of lexical and grammatical constituents' [Lexgr_12_diff] for the OHS_600 compounds. The results are highly significant ($p = 0.000$).

The spelling of compounds combining either two lexical or two grammatical constituents corresponds almost exactly to a chance distribution, with hyphenation only slightly below average (27.0 per cent). By contrast, compounds combining a lexical and a grammatical constituent in any order very clearly favour hyphenation (84.8 per cent), while dispreferring both solid spelling (9.1 per cent) and particularly open spelling (6.1 per cent). Hypothesis F8 is therefore supported by the data: compounds containing both lexical and grammatical constituents disfavour solid spelling (even if the results for open spelling are yet more extreme).

Table 5.54 *Combination of lexical and grammatical constituents [Lexgr_12_diff] in OHS_600*

			Combination of lexical and grammatical constituents		Total
			–	+	
OHS	o	Count	196	4	200
		Expected Count	178.0	22.0	200.0
		% within Lexgr_12_diff	36.7%	6.1%	33.3%
	h	Count	144	56	200
		Expected Count	178.0	22.0	200.0
		% within Lexgr_12_diff	27.0%	84.8%	33.3%
	s	Count	194	6	200
		Expected Count	178.0	22.0	200.0
		% within Lexgr_12_diff	36.3%	9.1%	33.3%
Total	Count		534	66	600
	Expected Count		534.0	66.0	600.0
	% within Lexgr_12_diff		100.0%	100.0%	100.0%

5.6.2 Syntactic Context

The influence of syntactic context on English compound spelling is discussed in many reference works. Merriam-Webster (2001: 107) advocates the hyphenation of adjective compounds in attributive position before a noun (***step-by-step** instructions*) but not in predicative position after the verb (*instructions that guide you **step by step***),¹⁶ when they function as subject complements or object complements (Quirk et al. 1985: 402–403) with open spelling. Excepted from this rule are two types of adjective compound which are supposedly hyphenated even in predicative position: firstly, compounds which continue “to function as a unit modifier” (e.g. *hikers who were **ill-advised** to cross the glacier* and *industries that could be called **low-tech***), and secondly, “[p]ermanent compound adjectives” (e.g. *for reasons that are **well-known***; cf. Merriam-Webster 2001: 107–108). Since both types of exception are difficult to operationalise, only the general spelling principle was subjected to testing in the form of two separate hypotheses (which are discussed jointly following the presentation of the individual results).

¹⁶ One may disagree with Merriam-Webster’s (2001) classification of *step+by+step* in the second example, since it is on a gradient to phrasal status. If classified as a compound, it should be considered an adverb rather than an adjective.

5.6.2.1 Hypothesis F9 – Attributive Position

Compounds in attributive position favour hyphenation. →
Tentatively confirmed.

Open spelling is possible in predicative position (*the clothing was **much needed***), because there is no danger of misunderstanding – in contrast to attributive use, where this is easily the case: thus *much-needed clothing*, without the hyphen, may be misinterpreted as *much* + [*needed clothing*] (Reiser 2007). The first part of the context-dependent spelling principle was therefore formulated as Hypothesis F9 (for all parts of speech):

F9: Compounds in attributive position favour hyphenation.

In order to test this hypothesis, all OHS_600 COMPOUNDS were searched in the written component of the British National Corpus (BNCwritten) with open, hyphenated and solid spelling in attributive position directly preceding a **noun**, e.g. LOW-TECH **variation** or COMPACT DISC **player**.¹⁷ In spite of the fact that some attributive uses could not be retrieved in this way (e.g. *a LOW-TECH early-Eighties video game* [CGC 2063] with the intervening sequence *early-Eighties*), a manual check of the corpus data suggests that the clear majority of instances should be found. The corpus frequencies with open, hyphenated and solid spelling were compared and the variant (o, h, s) occurring most frequently for each compound was coded in a separate column [Attr]. The code x marks the small number of cases with a draw between two or more spelling variants in BNCwritten.

Pearson's chi-square test for the dependent variable 'compound spelling in dictionaries' [OHS] and the independent variable 'preferred spelling in attributive position' is highly significant ($p = 0.000$) for the OHS_600 compounds. This is due to the fact that the dictionary spelling of the compounds usually coincides with the spelling in attributive position in BNCwritten: 82.0 per cent of the hyphenated OHS_600 compounds are hyphenated in attributive position, 79.5 per cent of the solid OHS_600 compounds are solid in attributive position and 51.5 per cent of the open OHS_600 compounds are spelled open in that syntactic context in BNCwritten. If we disregard the compounds without any corpus hits (e.g. an impressive 30.0 per cent of the open compounds), the proportions

¹⁷ Since only hyphenated and solid forms are lemmatised in the BNC (cf. 5.10.3), the use of the compounds' base forms made the results more comparable. The regular expressions for manual search in BNCweb (<http://bncweb.lancs.ac.uk>) would be *search word _{N}*, *search-word _{N}* and *searchword _{N}*, but the frequency data for the present study were automatically extracted from BNCwritten by means of a Perl script which does not search across sentence boundaries (in contrast to BNCweb).

Table 5.55 *Preferred spelling in attributive position in BNCwritten [Attr] and spelling in OHS_600*

			Dictionary spelling			
			o	h	s	Total
Spelling in attributive position	no corpus hits	Count	60	12	24	96
		Expected Count	32.0	32.0	32.0	96.0
		% within OHS	30.0%	6.0%	12.0%	16.0%
	open	Count	103	23	5	131
		Expected Count	43.7	43.7	43.7	131.0
		% within OHS	51.5%	11.5%	2.5%	21.8%
	hyphenated	Count	22	164	5	191
		Expected Count	63.7	63.7	63.7	191.0
		% within OHS	11.0%	82.0%	2.5%	31.8%
	solid	Count	5	0	159	164
		Expected Count	54.7	54.7	54.7	164.0
		% within OHS	2.5%	0.0%	79.5%	27.3%
	x (unclear)	Count	10	1	7	18
		Expected Count	6.0	6.0	6.0	18.0
		% within OHS	5.0%	0.5%	3.5%	3.0%
Total	Count	200	200	200	600	
	Expected Count	200.0	200.0	200.0	600.0	
	% within OHS	100.0%	100.0%	100.0%	100.0%	

even rise to 73.6 per cent coincidence for the open compounds, 87.2 per cent for the hyphenated compounds and 90.3 per cent for the solid compounds.

As regards the expected preference for hyphenation in attributive position in BNCwritten, there are indeed more hyphenations (191) than open (131) or solid spellings (164). However, only five compounds which are always solid in OHS_600 and only twenty-two compounds which are always open in OHS_600 change their usual spelling pattern and are preferably hyphenated in attributive position in BNCwritten (e.g. *tripwire* preceding *incident*; *time zone* preceding *transitions*). Twenty-three hyphe-nated OHS_600 compounds unexpectedly seem to prefer open spelling in attributive use (e.g. *drive-in* or *by-product*), but a manual check of the BNCwritten data reveals that this is due to many incorrect search results which are only superficially attributive compounds (e.g. *I am uncertain as to whether I am capable of laying a **drive in** concrete and after standardizing **by product** sector*), and which had better be ignored in the discussion. The corpus hits for open spelling therefore need to be handled with extreme

caution without very detailed manual post-editing. At the same time, this result suggests that the influence of syntactic context on compound spelling very probably adds to the confusion of inattentive readers, who simply register that a compound such as *ice cream* (which is always spelled open in OHS_extra) has an alternative hyphenated spelling *ice-cream*, without considering that this could be due to its occurrence in front of a noun (e.g. in *ice-cream soda*, which is always spelled with a hyphen and a blank in the dictionaries – but does not occur in BNCwritten). The results for Hypothesis F9 are therefore somewhat inconclusive, but in view of the slight tendency for compounds in attributive position to favour hyphenation, the hypothesis can be tentatively accepted: compounds in attributive position favour hyphenation.

5.6.2.2 Hypothesis F10 – Predicative Position

Compounds in predicative position favour open spelling. →
Tentatively confirmed.

The second part of the context-dependent spelling principle was formulated as Hypothesis F10:

F10: Compounds in predicative position favour open spelling.

By analogy to Hypothesis F9, all OHS_600 COMPOUNDS were searched in BNCwritten with open, hyphenated and solid spelling in predicative position directly following a **verb**, e.g. *he is* ABSENT-MINDED.¹⁸ A separate column coded the variant (o, h, s) occurring most frequently in predicative position [Pred] for each compound. However, the manual check of several compounds in BNCwritten revealed that a certain proportion of predicative uses occurred in contexts which do not directly follow the verb, e.g.

- ... *the space movies of the 1950s were quite* LOW-TECH [FB8 462], which would require the addition of an optional slot for an adverbial premodifier.
- *Indeed, Grimshaw strove to make it* LOW-TECH. [AKS 250], in which the adjective *low-tech* is an object complement (as in Quirk et al.'s 1985 definition of *predicative*). A regular expression intended to take more of

¹⁸ Since only hyphenated and solid forms are lemmatised in the BNC (cf. 5.10.3), the use of the compounds' base forms made the results more comparable. The regular expressions for manual search in BNCweb (<http://bncweb.lancs.ac.uk>) would be `_ {V} search word`, `_ {V} search-word` and `_ {V} searchword`, but the frequency data for the present study were automatically extracted from BNCwritten by means of a Perl script which does not search across sentence boundaries (in contrast to BNCweb).

Table 5.56 *Preferred spelling in predicative position in BNCwritten [Pred] and spelling in OHS_600*

			Dictionary spelling			
			o	h	s	Total
Spelling in predicative position	no corpus hits	Count	116	34	89	239
		Expected Count	79.7	79.7	79.7	239.0
		% within OHS	58.0%	17.0%	44.5%	39.8%
	open	Count	70	52	8	130
		Expected Count	43.3	43.3	43.3	130.0
		% within OHS	35.0%	26.0%	4.0%	21.7%
	hyphenated	Count	8	103	3	114
		Expected Count	38.0	38.0	38.0	114.0
		% within OHS	4.0%	51.5%	1.5%	19.0%
	solid	Count	1	0	92	93
		Expected Count	31.0	31.0	31.0	93.0
		% within OHS	0.5%	0.0%	46.0%	15.5%
	x (unclear)	Count	5	11	8	24
		Expected Count	8.0	8.0	8.0	24.0
		% within OHS	2.5%	5.5%	4.0%	4.0%
Total	Count		200	200	200	600
	Expected Count		200.0	200.0	200.0	600.0
	% within OHS		100.0%	100.0%	100.0%	100.0%

these cases into account would have to consider the various structures of possible noun phrases acting as objects (e.g. optional determiner + optional adjective + noun).

- *Barney **was** a super person and brilliant at his scientist job, but otherwise a trifle ABSENT-MINDED and generally helpless.* [JYF 12], in which the adjective *absent-minded* is only part of a subject complement which is so complex that its systematic inclusion in a regular expression makes no sense.

In order not to miss such uses of the compounds, **non-attributive** use [Nonattr] was calculated as an additional variable by subtracting the number of attributive occurrences (preceding a noun) from the total number of occurrences of each compound in all three spelling variants.

Pearson's chi-square test for the dependent variable 'compound spelling in dictionaries' [OHS] and the independent variable 'preferred spelling in predicative position' [Pred] was highly significant ($p = 0.000$). Once again (cf. 5.6.2.1), the most frequent spelling variant in the corpus usually coincides with the unanimous dictionary spelling: if

the many instances with no predicative hits in BNCwritten (239 = 39.8 per cent) are ignored, 83.3 per cent of the open OHS_600 compounds are spelled open in predicative position, 62.0 per cent of the hyphenated OHS_600 compounds are hyphenated in predicative position and 82.9 per cent of the solid OHS_600 compounds are spelled solid in that syntactic context.

As regards the expected preference for open spelling in predicative position in BNCwritten, there are indeed more open spellings (130) than hyphenations (114) or solid (93) spellings. Fifty-two compounds which are always hyphenated in OHS_600 and eight compounds which are always solid in OHS_600 change their usual spelling pattern and favour open spelling in predicative position in BNCwritten. This seems to suggest that Hypothesis F10 can be accepted and that compounds in predicative position favour open spelling. However, the manual check of a subset of the results revealed that a relatively large proportion of the corpus hits with open spelling is incorrect (e.g. *The new unified code has continued to influence **modern day** legal systems*, in which *modern day* following the verb *influence* is superficially predicative but should actually be classified as attributive, because it is structurally closer to the following compound noun *legal systems*). Hypothesis F10 can therefore only tentatively be confirmed without very detailed manual corpus analyses, opening up possibilities for future research.

In addition, Pearson's chi-square test was carried out for an alternative position-dependent measure (cf. earlier in this chapter), namely the independent variable 'preferred spelling in non-attributive position' [Nonattr] and the dependent variable 'compound spelling in dictionaries' [OHS] for the OHS_600 compounds. The results are highly significant ($p = 0.000$) and provide converging evidence for the correctness of Hypothesis F10 (cf. Table 5.57): once again, the unanimous dictionary spelling tends to coincide with the most frequent spelling variant in the corpus, and most corpus hits (256) use open spelling.

If we compare the spelling favoured by the OHS_600 compounds in attributive, predicative and non-attributive position (cf. Table 5.58), we find a majority of hyphenation in attributive use and a majority of open spelling both in predicative and non-attributive position, as expected. This is in line with Bolinger's (1967: 9–10) observation that general qualities are frequently expressed attributively (*a handy tool*), whereas transitory qualities are commonly expressed predicatively (*Are your tools handy?* = 'conveniently available'), and with the tendency of the attributive slot to call for a premodifying compound (Swan 2005: 359) and the tendency for the

Table 5.57 Preferred spelling in non-attributive position in BNCwritten [Nonattr] and spelling in OHS_600

			Dictionary spelling			Total
			o	h	s	
Spelling in non-attributive position	no corpus hits	Count	9	3	3	15
		Expected Count	5.0	5.0	5.0	15.0
		% within OHS	4.5%	1.5%	1.5%	2.5%
	open	Count	185	59	12	256
		Expected Count	85.3	85.3	85.3	256.0
		% within OHS	92.5%	29.5%	6.0%	42.7%
	hyphenated	Count	1	132	7	140
		Expected Count	46.7	46.7	46.7	140.0
		% within OHS	0.5%	66.0%	3.5%	23.3%
	solid	Count	2	0	174	176
		Expected Count	58.7	58.7	58.7	176.0
		% within OHS	1.0%	0.0%	87.0%	29.3%
	x (unclear)	Count	3	6	4	13
		Expected Count	4.3	4.3	4.3	13.0
		% within OHS	1.5%	3.0%	2.0%	2.2%
Total			Count	200	200	600
			Expected Count	200.0	200.0	600.0
			% within OHS	100.0%	100.0%	100.0%

Table 5.58 Spelling of the OHS_600 compounds favoured in attributive, predicative and non-attributive position in BNCwritten

	Attr	Pred	Nonattr
no corpus hits	96	239	15
open	131	130	256
hyphenated	191	114	140
solid	164	93	176
unclear	18	24	13

predicative slot to call for a postmodifying phrase: cf. *a five-year-old **child*** (attributive compound) but **a five years old **child*** (unacceptable attributive phrase), *the child **is** five years old* (predicative phrase) but not **the child **is** five-year-old* (unacceptable predicative compound due to lacking determiner). As a consequence, it appears that the recommendation of the style guides tested in Hypotheses F9 and F10 is not a purely random prescriptive

convention but is justified by more general patterns in the English language (even if most compounds in the data were spelled identically regardless of their syntactic context).

5.6.3 Compound-Internal Syntactic Structure

Since compounds condense propositional meaning into a very brief structure, the relationship between their constituents is only expressed implicitly, and various approaches attempt to identify the grammatical relationship holding between compound constituents by using clausal paraphrases (e.g. Marchand 1969). Quirk et al. (1985: 1570–1578) categorise noun and adjective compounds in terms of a paraphrasing clause structure consisting of the constituents subject, verb, object, complement and adverbial (with a subclassification of the last category into relations such as *place*, *time*, *instrumental* and *other*). This approach reveals substantial differences between superficially similar compounds such as *glow+worm* and *punch+card*, which both consist of a verb followed by a noun: while “The worm [S] glows [V]” can be described as a subject-verb structure, “X punches [V] the card [O]” corresponds to a verb-object construction (Quirk et al. 1985: 1570–1578). However, the analysis of compounds as reduced sentences is not without its problems, since it depends on a number of assumptions (Stein 1974: 321):

- a) the underlying sentences must be declarative active main sentences;
- b) they do not in themselves contain any derived or compounded forms;
- c) inflections for tense, number etc. do not appear in word formations.

The analysis is particularly problematic for compounds without a verbal element (e.g. *girlfriend* or *windmill*), which needs to be inferred for the sentential paraphrase (Stein 1974: 321). Thus *bird cage* with its competing possible sentential paraphrases poses the danger of subjective interpretation:

Table 5.59 *Sentential paraphrases of the compound bird+cage*
(following Stein 1974: 321)

Sentential paraphrase	Syntactic function of <i>bird</i>	Syntactic function of <i>cage</i>
a) <i>The cage is for the bird.</i>	complement of purpose	subject
b) <i>The bird lives in the cage.</i>	subject	complement of place
c) <i>X keeps the bird in the cage.</i>	object	complement of place

A pilot study investigated the potential of underlying clause structure as a determinant of English compound spelling by counting the frequency of open, hyphenated and solid (indented) example compound words for the thirty-one subcategories of syntactic paraphrase in Quirk et al. (1985: 1570–1578). The spelling of these examples can be assumed to have been incidental in the writing of the grammar, but their representativeness for the subgroups (which is assumed to extend to their spelling as well) makes them ideal testing material. A clear spelling preference or dispreference of about two-thirds or more was marked in the data (cf. Table 5.60 for selected examples). While the shallow categorisation (e.g. VO or SV) yielded no interesting results, the consideration of additional factors such as morphology (e.g. the presence of particular affixes in column 4) or the order of the constituents (cf. *fault-finding* vs. *chewing gum*) correlated with compound spelling. Since the crucial additional variables were, however, already covered by other hypotheses, the analysis of sentential paraphrases was not pursued any further.

5.6.4 Summary

Section 5.6 investigates grammar-related variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Part of speech of the compound
- Part of speech of the individual constituents
- Part-of-speech class of the compound (open/lexical vs. closed/grammatical)
- Part-of-speech class of the constituents (open/lexical vs. closed/grammatical)
- Frequency of the compound in attributive position in BNCwritten
- Frequency of the compound in predicative position in BNCwritten
- Frequency of the compound in non-attributive position in BNCwritten.

Statistical testing revealed a strong effect of several independent variables on the dependent variable ‘compound spelling’ [OHS]. The following hypotheses were confirmed:

- The part of speech of the compound influences its spelling. [F1]
- Adjective compounds favour hyphenation. [F2]

Table 5.60 *Syntactic structures of compounds following Quirk et al. (1985: 1570–1578)*

PoS	Example word	Clause type	Syntactic paraphrase	Sentential analogue	Preferred spelling		
					Proportion	Favoured	% favoured
n	'SUN, RISE	SV	subject+deverbal noun	The sun rises.	open 1 hyphenated 1 solid 10	S	83
n	'FAULT-, FINDING	VO	object+verbal noun in <i>-ing</i>	X finds faults.	open 0 hyphenated 7 solid 4	H -O	64
n	'TAX-, PAYER	VO	object+agential noun in <i>-er</i>	X pays tax(es).	open 3 hyphenated 4 solid 4	–	–
n	'PUNCH, CARD	VO	verb+object	X punches the card.	open 1 hyphenated 2 solid 4	-O	–
n	'CHEWING, GUM	VO	verbal noun in <i>-ing</i> +object	X chews gum.	open 7 hyphenated 1 solid 0	O -S	88
n	'WIND, MILL	SO	n+n (n ₁ powers/operates n ₂)	The wind powers the mill.	open 6 hyphenated 1 solid 1	O	75
n	'GIRL, FRIEND	SC	n+n (n ₂ is n ₁)	The friend is a girl.	open 6 hyphenated 1 solid 1	O	75
n	'SNOW, FLAKE	SC	n+n (n ₂ is/consists of n ₁)	The flake consists of snow.	open 3 hyphenated 0 solid 3	-H	–
adj	'OCEAN-, GOING	VA	adverbial+ <i>-ing</i> participle	X goes across oceans.	open 0 hyphenated 3 solid 0	H	100
adj	'GREY-, GREEN	verbless	adj+adj in a coordinating relation but with the focus slightly more on the second element	X is green but with a greyish tint.	open 0 hyphenated 12 solid 1	H -O	92

- The parts of speech of the constituents influence the compounds' spelling. [F7]
- The combination of lexical and grammatical constituents disfavours solid spelling. [F8]

For some hypotheses, the available data were too limited to permit any statistically significant conclusions, but it was still possible to observe certain tendencies:

- Adjective compounds with a compound-initial adverb ending in *-ly* tend to favour open spelling. [F3]
- Adjective compounds with a compound-initial adverb ending in *-ly* tend to disfavour solid spelling. [F3]
- Verb compounds tend to disfavour open spelling. [F4]
- Adverb compounds tend to disfavour open spelling. [F5]
- Contrary to expectations, grammatical compounds do not disfavour open spelling. [F6]

The following findings can only tentatively be accepted without intense manual post-editing of the corpus data to exclude unwanted hits:

- The results suggest that compounds which are spelled solid in dictionaries hardly ever vary their spelling, regardless of syntactic context. [F9 and F10]
- Compounds that are spelled open or hyphenated in dictionaries have a slightly higher (but still very small) tendency to be hyphenated before a noun and to be spelled open in other contexts. [F9 and F10]

Some of the results permit even more precise conclusions than e.g. the expected avoidance of one particular spelling for a number of variables:

- A compound-initial adjective favours open spelling and disfavours solid spelling. [F7]
- A compound-initial adverb favours hyphenation. [F7]
- A compound-initial grammatical word favours hyphenation. [F7]
- A compound-initial noun disfavours hyphenation. [F7]
- A compound-initial verb favours hyphenation. [F7]
- A compound-final adjective favours hyphenation. [F7]
- A compound-final grammatical word favours hyphenation. [F7]
- A compound-final noun disfavours hyphenation. [F7]
- A compound-final verb or adverb favours hyphenation. [F7]
- The combination of lexical and grammatical constituents favours hyphenation and disfavours open spelling. [F8]

In addition to the results for the hypotheses under consideration, another finding emerged from the detailed analysis of the material:

- Noun compounds disfavour hyphenation. [F1]

5.7 Semantics

The following sections discuss and investigate the influence of various semantic variables on the spelling of English compounds, namely the semantics of the whole compound and its constituents, the semantic relation between the constituents and the compound's degree of idiomaticity.

5.7.1 *Semantics of the Compound and Its Constituents*

The literature on English compound spelling notes that certain compounds with a particular meaning favour one spelling variant, e.g.:

- Merriam-Webster (2001: 101) hyphenates noun compounds designating units of measurement (e.g. *light-year* or *kilowatt-hour*).
- Adjective compounds referring to colour are almost exclusively hyphenated if each constituent can function as a noun (e.g. *red-orange fabric*, *the fabric was red-orange*) in contrast to the “usually unhyphenated” colour names whose first element can only be an adjective (e.g. *a bright red tie*; cf. Merriam-Webster 2001: 108–109).
- The *GPO Style Manual* (2008: 82) has a spelling rule for military titles as a highly specialised semantic category: “Do not hyphenate a civil or military title denoting a single office, but print a double title with a hyphen” (e.g. *commander in chief* and *major general* vs. *secretary-treasurer*).

In view of the extremely large and varied number of potential semantic categories of the whole compound and its constituents, these were not tested as potential determinants of spelling variant selection – with the exception of one particular type of constituent in Hypothesis G1.

5.7.1.1 *Hypothesis G1 – Second Constituent with General Reference*

Compounds ending in a constituent with general reference (e.g. *man*)
favour solid spelling. → Confirmed.

Some constituents in English word formation are situated on a gradient between lexical suffix and free constituent. The special status of so-called

affixoids is based on their characteristic of forming many compounds and of having relatively general semantics (Bußmann 2002 s.v. *Affixoid*, *Suffixoid*), e.g. *man* in *policeman*, *fireman* etc. In these contexts, *man* has a meaning comparable to that of the suffix *-er* and is characterised by a reduction of the vowel to *schwa* (as against the full vowel /æ/ in other contexts). Since compounds containing affixoids are always spelled solid – which is a general property of suffixations (Ritter 2005a: 56) but not necessarily of compounding – we may formulate the following hypothesis:

G1: Compounds ending in a constituent with general reference (e.g. *man*, *place*, *thing*) favour solid spelling.

In order to test Hypothesis G1, compounds containing the general nouns listed in Halliday and Hasan (1976: 274) – *people*, *person*, *man*, *woman*, *child*, *boy*, *girl*, *creature*, *thing*, *object*, *stuff*, *business*, *affair*, *matter*, *move*, *place*, *question* and *idea* – were marked in a separate column of the database. Plurals (e.g. *children*) and genitives (e.g. *man's*) were accepted in the coding process, but derivatives (e.g. *childhood*) were not. A further distinction was made between general nouns as the first constituent (*gen_1*, e.g. *place+mat*) and as the second constituent (*gen_2*, e.g. *birth+place*).

Twenty of the OHS_600 compounds contain a noun with general reference: three as the first constituent and seventeen as the second constituent. Since the small number of compounds with an initial general reference noun did not meet the requirements of statistical testing, the data were recoded as a binary distinction between general reference nouns in compound-final position and all other cases [General_n_r]. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped general reference nouns' [General_n_r] yielded statistically valid and highly significant results ($p = 0.001$). Compared to the unmarked cases, which have almost the expected distribution of 33.3 per cent for each spelling variant, compounds ending in general reference nouns clearly favour solid spelling (thirteen compounds = 76.5 per cent). Hypothesis G1 is thus confirmed by the data for the OHS_600 sample: compounds ending in a constituent with general reference favour solid spelling.

In the next step, the analysis was extended to the OHS_extra compounds (which comprise twenty-four compounds with a general reference noun as the first and ninety-two as the second constituent) to investigate position-dependence. Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'general

Table 5.61 *Grouped general reference nouns [General_n_r] and spelling in OHS_600*

			General reference noun		Total
			–	+	
OHS	o	Count	197	3	200
		Expected Count	194.3	5.7	200.0
		% within General_n_r	33.8%	17.6%	33.3%
	h	Count	199	1	200
		Expected Count	194.3	5.7	200.0
		% within General_n_r	34.1%	5.9%	33.3%
	s	Count	187	13	200
		Expected Count	194.3	5.7	200.0
		% within General_n_r	32.1%	76.5%	33.3%
Total	Count		583	17	600
	Expected Count		583.0	17.0	600.0
	% within General_n_r		100.0%	100.0%	100.0%

Table 5.62 *General reference nouns [General_n] and spelling in OHS_extra*

			General reference noun			Total
			–	First constituent	Second constituent	
OHS_extra	o	Count	2,273	13	31	2,317
		Expected Count	2,247.4	14.4	55.2	2,317.0
		% within General_n	60.6%	54.2%	33.7%	60.0%
	h	Count	324	1	0	325
		Expected Count	315.2	2.0	7.7	325.0
		% within General_n	8.6%	4.2%	0.0%	8.4%
	s	Count	1,151	10	61	1,222
		Expected Count	1,185.3	7.6	29.1	1,222.0
		% within General_n	30.7%	41.7%	66.3%	31.6%
Total	Count		3748	24	92	3864
	Expected Count		3,748.0	24.0	92.0	3,864.0
	% within General_n		100.0%	100.0%	100.0%	100.0%

reference noun' [General_n] yielded highly significant ($p = 0.000$) results for the OHS_extra compounds.

The results confirm the findings for OHS_600 compounds with final general reference nouns, that solid spellings are more frequent than would be expected by chance (61 instead of 29.1). By contrast, compounds with

Table 5.63 *Compound-final general reference nouns*
[General_n] and spelling in OHS_extra

	O	H	S
<i>affair</i>	1	0	0
<i>business</i>	1	0	0
<i>boy</i>	6	0	7
<i>child</i>	1	0	3
<i>girl</i>	3	0	2
<i>man</i>	6	0	44
<i>matter</i>	2	0	0
<i>object</i>	3	0	0
<i>people</i>	2	0	0
<i>person</i>	2	0	0
<i>place</i>	1	0	4
<i>stuff</i>	0	0	1
<i>woman</i>	3	0	0
TOTAL	31	0	61

initial general reference nouns behave very similarly to the unmarked cases. The OHS_extra analysis also reveals that a surprisingly large number of general reference constituents only occur in open compounds (cf. Table 5.63): *affair* (*love affair*), *business* (*show business*), *matter* (*grey matter*), *object* (*sex object*), *people* (*boat people*), *person* (*displaced person*) and *woman* (*career woman*). While the majority of the general nouns are thus inconclusive or tend towards open spelling in their small number of occurrences, *man* dominates with fifty counts, of which forty-four are solid (possibly due to its shortness; cf. 5.2.7 and 5.2.8) and thus skews the results. As a consequence, Hypothesis G1 can only be accepted superficially – or for the compound-final general reference noun *man*.

5.7.2 *Semantic Relation between the Constituents*

By analogy to clause structure (cf. 5.6.3), the semantic relations between the constituents of compounds are only expressed implicitly (e.g. Schmid 2011: 209), so that the precise relationships between compound constituents are often difficult to determine (Bauer 1983: 204). While some instances are presumably idiosyncratic (e.g. *spaghetti Western*, ‘a film about American cowboys in the Wild West, especially one made in Europe by an Italian director’; cf. Bauer 1983: 204 and LDOCE), other semantic relationships

Table 5.64 *Semantic relations between compound constituents*

Relation/Code	Description	Example		
		O	H	S
cause 21	2 causes 1	<i>tear gas</i>		
cause 12	1 causes 2	<i>heat rash</i>		
ownership 12	1 has 2	<i>lemon peel</i>		
originator 21	2 makes 1	<i>silk worm</i>		
originator 12	1 makes 2			<i>firelight</i>
use 21	2 uses 1	<i>steam iron</i>		
use 12	1 uses 2			<i>handbrake</i>
identity 11	1 is 1		<i>salad-salad</i>	
identity 21	2 is 1	<i>soldier ant</i>		
identity 12	1 is 2	<i>child prodigy</i>		
simultaneity 12	X is 1 and 2 at the same time		<i>singer-songwriter</i>	
hyponymy 12	1 is part of category 2	<i>palm tree</i>		
location 21	2 is in/at etc. 1	<i>field mouse</i>		<i>schoolfellow</i>
location 12	1 is in/at etc. 2	<i>taxi stand</i>		
purpose 21	2 is for 1	<i>horse doctor</i>		
origin 21	2 is from 1	<i>olive oil</i>		
material 21	2 is made of 1	<i>paper bag</i>		
topic 21	2 is about 1	<i>abortion vote</i>		
resemblance 21	2 is like 1			<i>bellflower</i>
time 21	2 is during 1	<i>night watch</i>		
VO 21	X 2s 1			<i>skyscraper</i>
SV 21	2 is			<i>crybaby</i>
alternation 12	1 or 2		<i>yes-no</i>	
name 21	2 is named after 1	<i>Wellington boot</i>		

are presumably more basic and recur more frequently than others (Sauer 1985: 286), e.g. compounds “whose head has a human referent will probably indicate something about the occupation or identity of the person referred to” and “names of animals and plants are likely to carry indications of appearance, habitat or location” (Adams 2001: 86). Even though it would be difficult to devise an “exhaustive classification . . . into a reasonably small set of types” (Huddleston and Pullum 2002: 1647), different models have been proposed to describe the semantic relations holding between the constituents of compounds, e.g. by Levi (1978: 76–77) and Marchand (1960a: 15–17). Table 5.64 provides an overview of the semantic relations between compound constituents, drawing on these predecessors as well as on Koch and Marzo’s motivating relations (2007: 11–12) and

Plag's (2010: 266) semantic categories. The descriptions in column 2 refer to the first constituent as 1 and to the second as 2. The last three columns list typical examples from the sources with their original spelling.

In view of the subjectivity of semantic criteria, which makes their application to intersubjective spelling heuristics problematic, only a limited number of comparatively simple semantic relations were coded in OHS_600 in order to permit testing the influence of semantic relations between the constituents, namely compounds containing

- identical constituents, e.g. *salad+salad* (coded *r* for *reduplication*).
- (quasi-)synonymous constituents (*s*), in which the 'is a' relation works in both directions (e.g. *border+line*: 'a border is a line' and 'a line is a border').
- antonymous constituents (*a*), e.g. in the adjective *stop+go*.
- hyponymous constituents, which are linked by a 'kind of' relation (cf. Murphy 2010: 122–123). A distinction was made between genus-species compounds (*h1*, e.g. *number+one*: 'one is a number') and species-genus compounds (*h2*, e.g. *match+stick*: 'a match is a kind of stick'). The number in the code indicates the semantically larger unit. Note that the hyponymous constituent can frequently stand for the whole compound: a *match* is the same thing as a *matchstick*. Still, such cases need to be distinguished from compounds containing a constituent which can stand for the whole compound but in which the relation between the constituents themselves is not hyponymous, e.g. *black+smith*: while the noun *smith* can be used synonymously with *blacksmith* (cf. LDOCE), the adjective *black* and the noun *smith* stand in no hyponymous relation to each other. Such instances were coded *p* (for *pars pro toto* 'a part for the whole'), with the number indicating the position of the more general constituent (e.g. *p1* for *broken+down* and *p2* for *black+smith*).
- co-hyponymous constituents (*c*), e.g. *fox+hound*, but these were only coded for close relations (excluding e.g. *basket+ball*, because the inanimacy of both constituents was considered too general).
- meronymous constituents (*m*), which are related by a 'part of' relation, e.g. *muscle+man*. In contrast to Murphy (2010: 122–123), only obligatory parts were considered meronyms ('a man (necessarily) has muscles'). This approach excludes e.g. *motor+boat* and *text+book*, since a boat does not necessarily require a motor and picture books may do without text. Since meronymy is asymmetrical (cf. Murphy 2010: 122–123), the larger unit was indicated by the number in the codes *m1* (e.g. *touch+type*) and *m2* (e.g. *muscle+man*).

Table 5.65 *Special semantic relations between the constituents [Sem_relation] in OHS_600*

Code	Description	Example	n
r	identical constituents	<i>win+win</i>	2
s	(quasi-)synonymous constituents	<i>border+line</i>	1
a	antonymous constituents	<i>stop+go</i>	1
h1	hyponymous constituents (genus-species)	<i>number+one</i>	3
h2	hyponymous constituents (species-genus)	<i>match+stick</i>	11
p1	compounds containing non-hyponymous/ non-meronymous constituents, the first of which can stand for the whole compound	<i>broken+down</i>	9
p2	non-hyponymous/non-meronymous constituents, the second of which can stand for the whole compound	<i>black+smith</i>	11
c	co-hyponymous constituents	<i>fox+hound</i>	10
m1	meronymous constituents, the first of which is the larger unit	<i>touch+type</i>	9
m2	meronymous constituents, the second of which is the larger unit	<i>muscle+man</i>	5

While these semantic relations are part-of-speech independent, they were only coded if both constituents of a compound belonged to the same part of speech (except if grammatical words were involved or in the case of the exceptional category *pars pro toto*, where a semantic relation between the constituents and the whole compound was explored). The analysis yielded merely 62 OHS_600 compounds with special semantic relations [Sem_relation].

While the number of compounds was too small to permit statistically valid testing for the subcategories, the data in Table 5.66 were analysed in order to search for tendencies. Two clear results emerge even without statistical backing:

- Compounds with co-hyponymous constituents (e.g. *nurse+maid*) tend to favour solid spelling (with eight out of ten hits).
- Compounds with meronymous constituents, the first of which is the larger unit (e.g. *touch+type*: typing involves touching), favour solid spelling (with eight out of nine hits).

In addition, two specific hypotheses derived from the literature, which involve the semantic relation between the constituents, were tested.

Table 5.66 Special semantic relations between the constituents [Sem_relation] and spelling in OHS_600

[illegible]

5.7.2.1 Hypothesis G2 – Species-Genus Compounds

Species-genus compounds (e.g. *palm tree*) favour open spelling. → Tentatively refuted.

The constituents of some compounds are hyponymously contained within each other, so that their meaning can be paraphrased as ‘constituent 1 is a kind of constituent 2’, e.g. *birch+tree* (‘a birch is a kind of tree’). Bauer (1983: 95) uses open spelling for all such species-genus compounds (*cod fish*, *beech tree*, *puppy dog*, *palm tree* and *boy child*). Since the same is true of *bunny rabbit*, *kiwi fruit* and *koala bear* in Macmillan (2007), the following expectation was formulated:

G2: Species-genus compounds (e.g. *palm+tree*, *puppy+dog*) favour open spelling.

The coding procedure described earlier (cf. 5.7.2) yielded eleven species-genus compounds (*h2* in Table 5.66), of which eight are spelled solid (*brushwood*, *coastline*, *hemline*, *matchstick*, *railroad*, *sandalwood*, *shoreline*, *storehouse*) and only three, i.e. roughly one-third, are spelled open (*duffel coat*, *quad bike*, *race meeting*). Hypothesis G2 must therefore be refuted for this small (and statistically) invalid sample: species-genus compounds do not favour open spelling but rather tend towards solid spelling.

5.7.2.2 Hypothesis G3 – (Quasi-)Reduplication

Compounds containing identical or quasi-identical constituents (e.g. *fifty+fifty*) favour hyphenation. → Tentatively confirmed.

One of the most extreme semantic relations is the doubling of the constituents. Since both Morton Ball (1951: 9) and the *GPO Style Manual* (2008: 83) advocate the use of a hyphen in compounds formed of repetitive terms (e.g. the interjection *bye-bye* or the adjective *paw-paw*),¹⁹ this was formulated as Hypothesis G3:

G3: Compounds containing identical or quasi-identical constituents (e.g. *fifty+fifty*) favour hyphenation.

For the coding procedure, cf. Section 5.7.2. Two OHS_600 compounds contain identical constituents (*r* in Table 5.66): *fifty-fifty* and *win-win*, both

¹⁹ A subcategory of reduplicative compounds are ad hoc compounds on the borderline between phrase and compound, which emphasise prototypical features of the referent, e.g. the final lexeme of ‘I’ll make the tuna salad, and you make the SALAD–salad’ (Ghomeshi et al. 2004: 308). These can only be classified as compounds in the present model if they are uninflected or apply inflection only once (e.g. *kid-kids* but not *I’m not leaving-leaving*; cf. Ghomeshi et al. 2004: 322–323).

of which are hyphenated. While supporting Hypothesis G₃ (that compounds containing identical or quasi-identical constituents favour hyphenation), the extremely small number of items calls for the analysis of larger quantities of data. Still, the results contradict Kuperman and Bertram's (2013: 952) observation that stronger semantic similarity increased the likelihood of open spelling (which should be strongest in identical items from a logical perspective).

5.7.3 *Idiomaticity*

Another semantic relation which may play a role in the spelling of English compounds is that between the semantics of the constituents and the semantics of the whole compound, i.e. the compound's degree of idiomaticity.

5.7.3.1 *Hypothesis G₄ – Idiomatic Meaning of the Compound*

Compounds with an idiomatic meaning component disfavour open spelling. → Confirmed.

Numerous reference works regard compounds' idiomaticity as one of the strongest arguments in favour of either concatenation or hyphenation (cf. Morton Ball 1939 for an extensive collection of quotes). While nine of the fifteen metaphorical compounds listed in Schmid (2011: 140–141) use solid spelling, however, idiomaticity as such does not preclude open spelling: idioms in the strictest sense (e.g. the classic example *to kick the bucket*) are even defined as the combination of several orthographic words with a joint meaning that “cannot be predicted from the meanings of the words themselves” (Palmer 1981: 36), and phrasal verbs (e.g. *give up*) are always spaced even if their meaning is often not derivable from their components. This suggests that the avoidance of open spelling based on idiomaticity, which is so frequently postulated in the literature, is a special feature of compounds:

G₄: Compounds with an idiomatic meaning component disfavour open spelling.

In order to test Hypothesis G₄, compounds with an idiomatic component were marked in the OHS_600 database. Simplifying the model described in Sanchez (2008), three categories of idiomaticity were coded:

1. **Literal** (*l*): all constituents can be used with their usual meaning in a paraphrase of the compound's meaning, e.g. *day+light* ‘the light during the day’, *moth+eaten* ‘(partly) eaten by moths’.

2. **Idiomatic** (*i*): one constituent cannot be used with its usual meaning in a paraphrase of the compound's meaning, e.g. *voice+box* 'larynx', whose second constituent is regarded as metaphorical extension of *box* (which usually implies straight sides in a man-made receptacle).
3. **Highly idiomatic** (*h*): a paraphrase making use of all the constituents with their usual meaning is unable to capture the meaning of the compound, e.g. *golden parachute* ('part of a business person's contract which states that they will be paid a large amount of money if they lose their job, for example if the company is sold'; cf. LDOCE)

As is commonly the case with semantic classifications, many borderline cases might be resolved in different ways by different researchers. In compounds comprising very dissimilar literal vs. idiomatic meanings (e.g. *night+cap* 'a soft hat that people used to wear in bed' vs. 'an alcoholic drink that you have at the end of the evening, just before you go to bed'; cf. LDOCE), the literal meaning as recorded in LDOCE or the OED was analysed by definition. In the categorisation of individual compounds, a distinction had to be made between semantic features which can be regarded as part of the word-formation pattern and semantic features adding necessary information: since a paraphrase of the compound *chicken wire*, 'wire for chickens', does not describe its meaning, 'wire in **net** form for making chicken **fences**', precisely enough, this compound was classified as belonging into the category *h*. By contrast, it is argued here that the implicit 'be against' in *fire+wall* ('the wall is against fire') is general enough to permit a classification of this compound as literal (*l*). Since *house+trained* can be understood broadly even without reference to excrements, it was also classified as literal ('trained for being in the house'). By contrast, the meaning of a simple paraphrase of the compounds *white goods* is so wide that it incorrectly refers to many goods with white colour that are not white goods (i.e. 'equipment used in the house, e.g. washing machines and refrigerators'; LDOCE). *White goods* was therefore classified as *h*.

The clear majority (403) of the 600 OHS_600 compounds have a literal (*l*) interpretation. Eighty-one compounds are idiomatic (*i*) in the sense that at least one constituent cannot be used with its usual meaning in a paraphrase of the compound's meaning, and the remaining 116 compounds belong in the category 'highly idiomatic' (*h*), for which a paraphrase making use of all the constituents is still completely unable to capture the meaning of a compound (e.g. the verb *to dead+head* cannot be explained without

referring to the additional concepts ‘flowers’ and ‘removal’ in some way or another). The compounds in this category are frequently unheaded, e.g. *bell + bottoms* (i.e. trousers).

In order to test Hypothesis G4, Pearson’s chi-square test was carried out for OHS_600 with the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘idiomaticity’ [Idiom]. The results are highly significant ($p = 0.003$). As expected, compounds with a non-literal meaning component disfavour open spelling (with proportions of 21.0 per cent in the category *i* and 25.0 per cent in the category *h*). Hypothesis G4 can therefore be accepted: compounds with an idiomatic meaning component disfavour open spelling. Solid spelling (40.7 per cent) and hyphenation (38.3 per cent) are almost equally likely for compounds with a certain amount of idiomaticity (*i*), whereas the highly idiomatic compounds (*h*) clearly favour hyphenation (43.1 per cent). While this may seem surprising in view of Taft (2004) and Rakić’s (2009: 74) conclusion that “opaque or partly opaque compounds are accessed faster if they are spelled concatenated, since the decomposition route is less effective in that case”, the present study’s finding can be explained by part of speech: 117 (59 per cent) of the 197 non-literal compounds are adjectives (e.g. *pie+eyed*, *tongue+tied*), a part of speech which correlates very strongly with hyphenation (cf. 5.6.1.2) and even overrides semantic considerations.

Table 5.67 *Idiomaticity [Idiom] and spelling in OHS_600*

			Idiomaticity			Total
			l Literal	i Idiomatic	h Highly idiomatic	
OHS	o	Count	154	17	29	200
		Expected Count	134.3	27.0	38.7	200.0
		% within Idiom	38.2%	21.0%	25.0%	33.3%
	h	Count	119	31	50	200
		Expected Count	134.3	27.0	38.7	200.0
		% within Idiom	29.5%	38.3%	43.1%	33.3%
	s	Count	130	33	37	200
		Expected Count	134.3	27.0	38.7	200.0
		% within Idiom	32.3%	40.7%	31.9%	33.3%
Total	Count		403	81	116	600
	Expected Count		403.0	81.0	116.0	600.0
	% within Idiom		100.0%	100.0%	100.0%	100.0%

5.7.4 Summary

Section 5.7 investigates semantic variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- One or more general nouns as constituents
- Selected semantic relations between the constituents
- Degree of idiomaticity (i.e. no complete semantic compositionality) of the compound.

The following hypothesis was confirmed:

- Compounds with an idiomatic meaning component disfavour open spelling. [G4]

The following hypothesis is only superficially confirmed by the data:

- A compound-final general reference noun favours solid spelling. While this seems to confirm Hypothesis G1, the effect depends so much on one single noun (*man*) that the effect can also be explained by length (three letters/one syllable). [G1]

The results for some hypotheses are based on such a small number of affected compounds that they could be considered inconclusive, even if certain tendencies can be observed:

- Unexpectedly, species-genus compounds do not favour open spelling. [G2]
- As expected, compounds with identical constituents favour hyphenation. [G3]

In addition to the results for the hypotheses under consideration, other findings emerged from the detailed analysis of the material (but note that some of these are based on very small numbers):

- Compounds with co-hyponymous constituents favour solid spelling. [G]
- Compounds containing meronymous constituents, the first of which is the larger unit, favour solid spelling. [G]
- Highly idiomatic compounds favour hyphenation. [G4]

To conclude, the influence of semantic variables on the spelling of English compounds is less clear than that of the other variables investigated so far,

because many of the semantic relations investigated are very specific and therefore only apply to a very restricted number of compounds.

5.8 Diachronic Variables

Since the present-day shape of any word is also the result of its historical development, one may assume that diachronic variables like the age of the compound and the language of origin of the whole compound and of its constituents have some effect on the spelling.

5.8.1 *Language of Origin*

Foreign language of origin is commonly used as a reason for spelling compounds less closely than native compounds: thus the *GPO Style Manual* (2008: 79) demands not to hyphenate “a unit modifier consisting of a foreign phrase” (e.g. in *bona fide transaction* and *prima facie evidence*), but the unit status of foreign constructions functioning as a compound constituent is often indicated by italicisation (Quirk et al. 1985: 1635), whereas hyphenation is expected for native adjective lexemes (cf. 5.6.2). Merriam-Webster (2001: 106) exceptionally accepts hyphenation if adjective compounds “composed of foreign words” are hyphenated in their language of origin. The spelling conventions of the language from which complex loan constructions are borrowed may thus exert some influence on English compound spelling. Bauer (1998: 84) even suspects that “the fact that English happens to write some noun + noun collocations as one word and others as two” arises from “the fact that English is a Germanic language strongly influenced by a Romance one (namely French), and that the current situation is a blend of conventions from these two sources”.

What is problematic about these diachronic principles is the question what should be considered foreign in a language like English, which has an extremely strong borrowing tradition. Since a considerable number of high-frequency Romance words are perfectly naturalised and therefore barely recognisable as linguistic immigrants to the uninitiated (e.g. *family*, *state*, *add* or *pay*; cf. the frequency lists in Sanchez 2008), such diachronic principles are likely not to address merely the historical origin of words but implicitly also their recognisability as loans (cf. 5.8.1.2).

5.8.1.1 Hypothesis H1 – Combination of Germanic and Romance Constituents

Compounds containing one Germanic and one Romance constituent (e.g. *time+zone*) disfavour solid spelling. → Confirmed.

Foreign origin is a particularly interesting variable for compounds with mixed sources. If one explores the idea of the combination of heterogeneous constituents further (cf. earlier and 5.12.2), it entails that combining a native Germanic constituent with a constituent of Romance (or otherwise borrowed) origin might disfavour solid spelling:

H1: Compounds containing one Germanic and one Romance constituent (e.g. *time+zone*) disfavour solid spelling.

In order to test this hypothesis, the language of origin was coded for the constituents of the OHS_600 compounds based on the etymological information from the *Oxford English Dictionary*. Following the model developed in Sanchez (2008: 115–127), a distinction concerning the proximate language (cf. Hillebrand 1975: 224) was drawn between six categories:

- **Germanic** (*g*): used in a very wide sense and applied to all constituents which entered English via some Germanic language (e.g. Old Norse) or are already documented in Old English – even if they entered that earliest stage of English from other languages (*pepper*, *camp*) or if their etymology beyond Old English is doubtful (*narrow*)
- **Romance** (*r*): used in a very wide sense, comprising loans from the Romance languages (e.g. Latin and French) and words which entered English via the Romance languages (e.g. Greek loans like *system*, which were usually borrowed into English from Latin)
- **Mixed** (*m*): mixed Germanic-Romance origin, e.g. Germanic suffixations of Romance bases (*snake+charmer*) or Romance suffixations of Germanic bases (*cooker+book*). Inflectional suffixes (*human+resources*) and the borderline lexical suffixes *-ing/-ed* (*boarding+card*), all of which are Germanic (cf. Sanchez 2008: 138–139), were ignored
- **Onomatopoeic** (*o*): onomatopoeic constituents (*flick*)
- **Name** (*n*): names or constituents derived from names (*jeans*, *sandwich*)
- **Unclear** (*u*): unclear etymology in the OED (*trash*).

The codes for the constituents' language of origin were combined in one column of the database (e.g. *gr* for *day+return* and *rg* for *police+dog*; cf. Table 5.68), and mixed origin was also marked in a separate column

Table 5.68 *Etymological origin of the constituents [Etym_orig] of the OHS_600 compounds*

		Example	Frequency	Per cent
Code	gg	<i>black+smith</i>	279	46.5
	gm	<i>snake+charmer</i>	1	0.2
	gn	<i>house+martin</i>	1	0.2
	go	<i>chick+flick</i>	1	0.2
	gr	<i>time+zone</i>	92	15.3
	gu	<i>white+board</i>	12	2.0
	mg	<i>cookery+book</i>	3	0.5
	nr	<i>sandwich+course</i>	2	0.3
	nu	<i>duffel+bag</i>	1	0.2
	og	<i>shoo+in</i>	2	0.3
	or	<i>crash+helmet</i>	2	0.3
	ou	<i>jam+packed</i>	1	0.2
	rf	<i>fact+finding</i>	1	0.2
	rg	<i>false+friend</i>	91	15.2
	rn	<i>male+chauvinist</i>	1	0.2
	rr	<i>petrol+station</i>	79	13.2
	ru	<i>candy+floss</i>	10	1.7
	ug	<i>beach+head</i>	14	2.3
	ur	<i>wrapping+paper</i>	5	0.8
	uu	<i>baby+boom</i>	2	0.3
	TOTAL		600	100.0

[Mixed_etym]. The most frequent combination for the OHS_600 compounds is *Germanic+Germanic* (279), followed by 183 compounds combining constituents with Germanic and Romance origin (in any order) and *Romance+Romance* (79).

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'mixed Germanic and Romance etymological origin' [Mixed_etym] is significant ($p = 0.014$) for OHS_600. As expected, solid spelling is least likely among the compounds with mixed etymology. Hypothesis H1 can therefore be confirmed: solid spelling is disfavoured by compounds containing one Germanic and one Romance constituent (even if solid spelling is still relatively common with a proportion of 27.3 per cent). The most important finding emerging from Table 5.69 is that compounds with mixed etymology are characterised by a relatively strong tendency towards open spelling (41.5 per cent) – and not hyphenation, as one may have expected in view of the fact that these compounds combine heterogeneous parts.

Table 5.69 *Mixed Germanic and Romance origin [Mixed_etym] and spelling in OHS_600*

			Mixed etymological origin		Total
			-	+	
OHS	o	Count	124	76	200
		Expected Count	139.0	61.0	200.0
		% within Mixed_etym	29.7%	41.5%	33.3%
	h	Count	143	57	200
		Expected Count	139.0	61.0	200.0
		% within Mixed_etym	34.3%	31.1%	33.3%
	s	Count	150	50	200
		Expected Count	139.0	61.0	200.0
		% within Mixed_etym	36.0%	27.3%	33.3%
Total	Count		417	183	600
	Expected Count		417.0	183.0	600.0
	% within Mixed_etym		100.0%	100.0%	100.0%

Table 5.70 *Germanic/Romance/mixed origin of constituents, average length and spelling in OHS_600*

Origin of constituents	Spelling			No. of items	Average length (letters)
	O	H	S		
Germanic +	34	114	131	279	8.7
Germanic	12.2%	40.9%	47.0%		
Mixed (g+r / r+g)	76	57	50	183	9.9
	41.5%	31.1%	27.3%		
Romance +	64	11	4	79	11.9
Romance	81.0%	13.9%	5.1%		

Table 5.70 shows that compounds with one Germanic and one Romance constituent assume an intermediate position between purely Germanic and purely Romance compounds regarding the proportion of spellings for the three variants (e.g. 41.5 per cent open spellings in the mixed group as against 12.2 per cent for the Germanic and 81.0 per cent for the Romance compounds). However, with their general preference for open spelling and their slight dispreference of solid spelling, the etymologically mixed compounds behave more like purely Romance than like purely Germanic compounds.

5.8.1.2 *Hypothesis H2 – Synchronically Felt Foreign Origin*

Compounds giving the impression of being foreign words disfavour solid spelling. → Confirmed.

Many loanwords are so perfectly integrated into the English language that they are hard to recognise without prior linguistic training (cf. Section 5.8.1). Since, however, spelling principles addressing the foreign origin of words need to be related to their present-day recognisability as borrowings if they are to be used by non-linguists, ‘synchronically felt foreignness’ (cf. Munske 1983) was investigated as a potential determinant of English compound spelling:

H2: Compounds giving the impression of being foreign words disfavour solid spelling.

In order to objectivise this intrinsically subjective variable as much as possible, synchronically felt foreignness was coded in a separate column of the database if at least one constituent of an OHS_600 compound contained:

- unusual phonemes – e.g. nasalized /*æ̃*/ as a possible sound in the French-based English loanword *coq au vin* (cf. Roach, Hartman and Setter 2006 s.v. *coq au vin*)
- stress on non-initial syllables (e.g. in *patrol*) as against the Germanic tendency to stress the first syllable (cf. e.g. Meyer et al. 2005: 122; Celce-Murcia, Brinton and Goodwin 1996: 133–134)
- graphemes with unusual diacritics such as the circumflex in the French-based English loanword *bête noire* (LDOCE)
- unusual grapheme–phoneme correspondences – e.g. the silent word-final <n> in *coq au vin*)
- unusual combinations of graphemes or phonemes – e.g. the word-initial <schw> in Hebrew-based *schwa*
- (sequences of) graphemes in an unusual position, e.g. the constituent-final in *pistol*
- unusual morphology like the “foreign” plural forms *appendices*, *cacti* and *formulae* (cf. Swan 2005: 516–517)
- particularly long constituents, such as *correspondence*.

While synchronically felt foreignness need not necessarily coincide with actual etymological origin, all forty-seven affected OHS_600 compounds (e.g. *anabolic+steroid* or *systems+analyst*) were found to be loanwords.

Table 5.71 *Synchronically felt foreignness [Foreignness] and spelling in OHS_600*

			Synchronically felt foreignness		Total
			-	+	
OHS	o	Count	160	40	200
		Expected Count	184.3	15.7	200.0
		% within Foreignness	28.9%	85.1%	33.3%
	h	Count	195	5	200
		Expected Count	184.3	15.7	200.0
		% within Foreignness	35.3%	10.6%	33.3%
	s	Count	198	2	200
		Expected Count	184.3	15.7	200.0
		% within Foreignness	35.8%	4.3%	33.3%
	Total	Count	553	47	600
		Expected Count	553.0	47.0	600.0
		% within Foreignness	100.0%	100.0%	100.0%

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'synchronically felt foreignness' [Foreignness] is highly significant ($p = 0.000$) for the OHS_600 compounds. Compounds with synchronically felt foreignness clearly favour open spelling (85.1 per cent). Solid spellings are least likely in this group with a proportion of 4.3 per cent, closely followed by hyphenation. Hypothesis H2 can therefore be accepted: compounds giving the impression of being foreign words disfavour solid spelling.

5.8.2 Age of the Compound

A recurrent idea about the spelling of English compounds is that it develops from open to concatenated spelling via hyphenation as an intermediate step (cf. Chapter 1). If this were the case, one should expect a correlation between the age of individual compounds and their preferred spelling variant: recent compounds (which have had only little opportunity to change) should tend towards open spelling, whereas older compounds could be expected to prefer concatenation. Compounds that are neither old nor new would favour hyphenation in such a model. These assumptions were tested in the form of several individual hypotheses.

5.8.2.1 Hypothesis H₃ – Old Compounds

Old compounds favour solid spelling. → Confirmed.

The first assumption was formulated as hypothesis H₃:

H₃: Old compounds favour solid spelling.

In order to test this hypothesis, the age of the compounds was coded by using the *Oxford English Dictionary* as the most comprehensive etymological resource for English. For each OHS_600 compound, the date next to the oldest quotation (which corresponds to “its earliest recorded usage”; cf. OED CD-ROM, “Sense section”) was recorded, e.g. 1951 for *place+mat*. For some dates, slight modifications were made in a separate column:

- the abbreviation *a* (= *ante* ‘before’, e.g. *a1000* for *breast+bone*) was ignored – yielding the date 1000 in the example.
- the abbreviation *c* (= *circa*, e.g. *c1430* for *work+day*) was ignored – yielding the date 1430 in the example.
- the question mark preceding a date (only in the case of *light-headed ?1537*) was ignored.
- periods (e.g. *1727–52* for *milk+tooth*) were reduced to the initial date – yielding 1727 in the example.
- incomplete dates (e.g. *177.* for *labour+saving*) were completed by using a *0* instead of the full stop – yielding the date 1770 in the example.

The date was coded regardless of the precise meaning used in the first attestation (e.g. *aircraft* before the invention of planes), but only for the matching part of speech (e.g. for the adjective *drip+dry* rather than earlier verbal uses). Quotations which are completely in Latin (e.g. for *seal+skin*) and quotations in angled brackets (e.g. for *brown+stone*) were ignored and the subsequent quotation was used.

Table 5.72 gives an overview of the centuries in which the OHS_600 compounds are attested for the first time [Age]. More than half of the compounds were first used in the nineteenth or twentieth centuries, whereas the early centuries of the English language are clearly underrepresented.

In order to meet the requirements for statistical testing, the eleven compounds without an indication of age (e.g. *fire+hydrant* and *marital+status*) were ignored and the individual dates, which are precise to the year, were grouped. A distinction was then made between old compounds (with a first attestation up to the year 1500), recent compounds (dating from 1900 or later) and a third group comprising the

Table 5.72 *Date of first attestation of the OHS_600 compounds in the OED [Age] by century*

Century	Example	Frequency
no date	<i>banker's+order</i>	11
ninth	<i>shell+fish</i>	1
tenth	<i>god+son</i>	3
eleventh	<i>snow+white</i>	10
twelfth	<i>wild+fire</i>	2
thirteenth	<i>grind+stone</i>	6
fourteenth	<i>day+light</i>	18
fifteenth	<i>long+lived</i>	17
sixteenth	<i>farm+house</i>	58
seventeenth	<i>dragon+fly</i>	53
eighteenth	<i>school+girl</i>	51
nineteenth	<i>safety+valve</i>	174
twentieth	<i>user+friendly</i>	196

intermediate period. Furthermore, the OHS_600 compounds' spelling in the first OED attestation was coded. While the spelling in the OED may not always reflect the original spelling of the first attestation due to the varying editorial practices of its sources, the same problem occurs even with specialised historical dictionaries, which contain less information on compounds than the OED. The OED even uses a tilde in quotations to mark "a hyphen introduced in the printing of the First Edition of the *OED*, which may not have been present in the cited text" (OED CD-ROM; *Key to symbols and other conventions*). Where this was the case (e.g. for *birth+right*), the spelling of the next suitable example was recorded.

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped age of the compound' [Age_r] was highly significant ($p = 0.000$) for the 589 OHS_600 compounds with a date of first attestation in the *Oxford English Dictionary*. As predicted by Hypothesis H3, the fifty-nine old compounds clearly favour solid spelling (67.8 per cent). Both hyphenation (20.3 per cent) and open spelling (11.9 per cent) are clearly disfavoured. Hypothesis H3 can therefore be accepted: old compounds favour solid spelling.

Table 5.73 *Grouped age of the compound [Age_r] and spelling in OHS_600*

			Age of the compound			Total
			old (–1500)	intermediate (1501–1899)	new (1900–)	
OHS	o	Count	7	101	83	191
		Expected Count	19.1	108.3	63.6	191.0
		% within Age_r	11.9%	30.2%	42.3%	32.4%
	h	Count	12	113	73	198
		Expected Count	19.8	112.3	65.9	198.0
		% within Age_r	20.3%	33.8%	37.2%	33.6%
	s	Count	40	120	40	200
		Expected Count	20.0	113.4	66.6	200.0
		% within Age_r	67.8%	35.9%	20.4%	34.0%
	Total	Count	59	334	196	589
		Expected Count	59.0	334.0	196.0	589.0
		% within Age_r	100.0%	100.0%	100.0%	100.0%

5.8.2.2 *Hypothesis H4 – Recent Compounds*

Recent compounds disfavour solid spelling. → Confirmed.

The second hypothesis which was derived from the assumed development of compounds’ spelling through time (cf. 5.8.2) is the following:

H4: Recent compounds disfavour solid spelling.

Hypothesis H4 was tested simultaneously with Hypothesis H3 (cf. 5.8.2.1 for the methodology). Pearson’s chi-square test for the dependent variable ‘compound spelling’ [OHS] and the independent variable ‘grouped age of the compound’ [Age_r] was highly significant ($p = 0.000$) for the 589 OHS_600 compounds with a date of first attestation in the *Oxford English Dictionary*. The recent compounds behave in a way which is directly opposed to the results for the old compounds (except that the proportions are less extreme): there is a clear preference for open spelling (42.3 per cent); hyphenation comes second (37.2 per cent) and is followed by a clear dispreference for solid spelling (20.4 per cent). Hypothesis H4 is thus supported by the data: recent compounds disfavour solid spelling. Another interesting observation is that the compounds which are neither very old nor very recent behave almost as one would have expected in a chance distribution: their proportion of roughly one-third for each spelling variant reflects their diachronically unmarked status.

5.8.2.3 Hypothesis H₅ – Development from Open to Hyphenated to Solid Spelling

Compounds usually develop from open spelling via a hyphenated intermediate step towards solid spelling. → Refuted.

In addition to the tests carried out on particularly old and recent compounds (cf. 5.8.2.1 and 5.8.2.2) as the most extreme cases, the assumed development of compounds' spelling through time (cf. Chapter 1) was also tested:

H₅: Compounds usually develop from open spelling via a hyphenated intermediate step towards solid spelling.²⁰

If Hypothesis H₅ were true, we should expect to find that in the course of a specific period of time, a majority (i.e. more than 50 per cent) of the compounds under consideration have developed

- a) from open to hyphenated spelling (with the potential to move on to solid spelling in the future)
- b) from hyphenated to solid spelling (because their open stage was already finished at the beginning of the observation period)
- c) from open to solid spelling (because the hyphenated stage was completed within the observation period).

A preservation of the original spelling is also compatible with Hypothesis H₅ (because the observation period might be too short to observe any developments for particular compounds). By contrast, a predominance of developments

- a) from solid to open spelling
- b) from solid to hyphenated spelling
- c) from hyphenated to open spelling

would contradict the hypothesis. Hypothesis H₅ was tested on the OHS_600 compounds firstly, by drawing on historical information from the *Oxford English Dictionary*, and secondly, by carrying out a corpus study on the Brown family corpora.

The first test extended over the longest possible diachronic observation period for each compound by using the spelling in the first OED

²⁰ While the literature posits a development from open via hyphenated to solid spelling, the expectation of the present study was actually that this is not the norm, and that the longer a word has been in existence, the more systematic its spelling will become. Nonetheless, the more common hypothesis was used for testing.

Table 5.74 *Spelling development of the OHS_600 compounds from their first attestation in the OED to their unanimous present-day dictionary spelling*

Spelling development	Frequency	%
-	11	1.8
hh	143	23.8
ho	46	7.7
hs	80	13.3
oh	44	7.3
oo	144	24.0
os	72	12.0
sh	11	1.8
so	1	0.2
ss	48	8.0
TOTAL	600	100.0

attestation (cf. 5.8.2.1) as the starting point and the unanimous dictionary spelling in present-day dictionaries as the end point of potential developments in the spelling. Note that this method cannot determine whether the spelling fluctuated and possibly even reversed several times between the first attestation and present-day spelling. The results are summarised in Table 5.74.

In this approach, the observation period is roughly identical with the complete span of existence of the 589 OHS_600 compounds listed in the OED (nine open and two hyphenated compounds lacked information on first attestation). If Hypothesis H5 were true, one should therefore expect to find only (or at least mainly) open spellings in the first attestations. That is, however, not the case, as there are 260 open compared to 269 hyphenated and 60 solid items. The claim that compounds start out their life with open spelling can thus be refuted. This finding is in line with Bauer (1998: 69), who notes that some neologisms in the *Oxford Dictionary of New Words* (Tulloch 1991) are spelled as single orthographic words, without any evidence that they were ever spelled differently (e.g. *airside*), whereas some item-familiar constructions are still written with an intervening space (e.g. *college degree*). Another interesting observation from Table 5.74 is that the earliest attestations use solid spelling considerably less frequently than the other two variants. Of the compounds which were hyphenated right from the beginning, many are adjectives (114 of the 157 OHS_600 adjectives), but the majority of the hyphenations are still nouns

(144). This lends support to the view that newness is and has been frequently indicated by the use of hyphens (cf. 7.1).

Most importantly for the present study, Table 5.74 shows that 335 (55.8 per cent) of the OHS_600 compounds are still spelled now as they were upon first entering the English language: 143 of the 200 present-day hyphenations have always been hyphenated, and 144 of the 200 present-day open forms have always been spaced. These results are supported by statistical testing: Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'grouped earliest spelling in the *Oxford English Dictionary*' [Earl_spell_r] is highly significant ($p = 0.000$). This is due to the fact that

- initial open spelling favours present-day open spelling (55.4 per cent) and disfavors present-day hyphenation (16.9 per cent)
- initial hyphenated spelling favours present-day hyphenation (53.2 per cent) and disfavors present-day open spelling (17.1 per cent)
- initial solid spelling favours present-day solid spelling (80.0 per cent). Present-day open spelling (1.7 per cent) is even less likely than present-day hyphenation (18.3 per cent).

If we consider the data from a different perspective, merely forty-eight of the now exclusively solid compounds started with solid spelling, whereas seventy-two were first attested with open spelling and as many as eighty started as hyphenations. This suggests that solid spelling is indeed frequently the result of a development (in contrast to open spelling and hyphenation, which are already present in the majority of the first attestations of present-day open and hyphenated compounds). Where changes in the spelling occurred, those involving a stronger degree of concatenation dominate with 132 compounds as against 58 instances with the opposite tendency – i.e. one development from solid to open spelling (e.g. *video+camera*), eleven from solid to hyphenated spelling (eight adjectives and three nouns, e.g. *glow+worm*) and forty-six from hyphenated to open spelling (all nouns, e.g. *safety+razor*). In line with Haiman's (1983: 781) principles of iconic and economic motivation, which posit that "[t]he distance between linguistic expressions may be an iconically motivated index of the conceptual distance between the terms or events which they denote" and that "[r]educd form may thus be an economically motivated index of familiarity", we can conclude from these findings that strong orthographic links are unlikely to become weaker in the spelling development of compounds through time.

Table 5.75 *Grouped earliest spelling in the OED [Earl_spell_r] and spelling in OHS_600*

			Earliest spelling			Total
			open	hyphenated	solid	
OHS	o	Count	144	46	1	191
		Expected Count	84.3	87.2	19.5	191.0
		% within Earl_spell	55.4%	17.1%	1.7%	32.4%
	h	Count	44	143	11	198
		Expected Count	87.4	90.4	20.2	198.0
		% within Earl_spell	16.9%	53.2%	18.3%	33.6%
	s	Count	72	80	48	200
		Expected Count	88.3	91.3	20.4	200.0
		% within Earl_spell	27.7%	29.7%	80.0%	34.0%
Total	Count		260	269	60	589
	Expected Count		260.0	269.0	60.0	589.0
	% within Earl_spell		100.0%	100.0%	100.0%	100.0%

Taking all of this into account, a strict version of Hypothesis H5 must be refuted by the first test: compounds do not generally start with open spelling, then become hyphenated and then solid. However, if compounds were open or hyphenated upon their first occurrence, they had a chance of 40 per cent and 36 per cent, respectively, of becoming solid up to the present. Nonetheless, open and hyphenated spellings may be relatively stable: *black pepper* (the oldest open compound which conserves its form) dates back to about 1000; *wild boar* and *precious stone* are from the thirteenth century. The oldest hyphenation (*ice-cold*) dates from about 1000, and *moth-eaten* was first attested in the fourteenth century. While some might claim that these individual compounds are still on their way towards unavoidable solid spelling and simply require more time to complete the change, the alternative view adopted here rather regards open and hyphenated spellings as consistent choices for particular types of compound, which need not necessarily change in the future. Such a view is in line with Peters (2004: 119), who regards length and part of speech as preventing solid spelling: she assumes that longer compounds like *daylight-saving* may never become solid, regardless of how well established they are, whereas many compound adjectives or verbs will generally avoid open spelling, because hyphenation or solid spelling support their reading as single units.

Table 5.76 *Spelling of the OHS_600 compound types in chronologically ordered British English corpora*

	B-LOB (1931)	LOB (1961)	FLOB (1991)	BEo6	TOTAL
O	119 (39.5%)	125 (37.8%)	137 (38.3%)	155 (41.3%)	536
H	106 (35.2%)	115 (34.7%)	108 (30.2%)	99 (26.4%)	428
S	76 (25.2%)	91 (27.5%)	113 (31.6%)	121 (32.3%)	401
TOTAL	301	331	358	375	

The first test of Hypothesis H₅ used dictionary rather than corpus data for the largest possible observation period, since the earliest recorded use of compounds as disparate as those in the OHS_600 list could not be retrieved from any single corpus at the time of the present study. The analysis of spelling differences, capitalisation and punctuation from the OED's first attestations further justifies this procedure, as the original spelling of the constituents of 119 compounds (e.g. *hand+**boc***) differs from the currently used standard (*hand+**book***). This spelling variation, which would have prevented or greatly complicated retrieval, is also the reason why the corpus study as the second test of Hypothesis H₅ was restricted to relatively recent material, the British English Brown family corpora (cf. 4.2). Covering a span of merely seventy-five years, the period (1931–2006) is possibly too short to capture any major changes, but this complementary study can shed some light on more recent developments. The absolute numbers and percentages in Table 5.76 refer to the OHS_600 compound types with open, hyphenated and solid spelling found in the respective corpus. A particular compound may thus have contributed to all three counts if it occurred in all three spelling variants. If Hypothesis H₅ was correct, we should expect to find more open and/or hyphenated spellings in the earlier corpora and more hyphenated and/or solid spellings in the corpora of more recent English, since the same set of compounds was used for all searches.

A chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'time of corpus' was carried out using *Excel*'s CHITEST function for the data in Table 5.76. While the results are not statistically significant ($p = 0.115$), there is an obvious decrease in the proportion of hyphenations (from 35.2 per cent in 1931 to 26.4 per cent

Table 5.77 *Spelling of the OHS_600 compound types in chronologically ordered British English corpora, combined with the unanimous dictionary spelling*

Dictionary	Corpus	B-LOB (1931)	LOB (1961)	FLOB (1991)	BE06 (2006)
O	O	52 (81.3%)	61 (84.7%)	71 (81.6%)	87 (85.3%)
	H	12 (18.8%)	10 (13.9%)	13 (14.9%)	10 (9.8%)
	S	0 (0.0%)	1 (1.4%)	3 (3.4%)	5 (4.9%)
H	O	43 (37.7%)	44 (34.4%)	54 (35.8%)	57 (36.5%)
	H	65 (57.0%)	82 (64.1%)	86 (57.0%)	84 (53.8%)
	S	6 (5.3%)	2 (1.6%)	11 (7.3%)	15 (9.6%)
S	O	24 (19.5%)	20 (15.3%)	12 (10.0%)	11 (9.4%)
	H	29 (23.6%)	23 (17.6%)	9 (7.5%)	5 (4.3%)
	S	70 (56.9%)	88 (67.2%)	99 (82.5%)	101 (86.3%)

in 2006), whereas solid spellings are clearly on the increase (from 25.2 per cent in 1931 to 32.3 per cent in 2006) – all of which would seem to support Hypothesis H5. By contrast, the proportion of compounds with open spelling, which is always the largest category (and needs to be considered with caution due to superficially identical phrasal sequences; cf. 5.3.2), vacillates with a tendency towards increase.

Table 5.77 combines the corpus hits for each spelling variant in the British English corpora with the unanimous dictionary spelling of each compound. Since this is a more detailed account of the data in Table 5.76, no additional test of statistical significance was carried out. The percentages refer to the distribution of corpus results for each dictionary-based spelling variant.

The shaded cells indicate that for each spelling variant, the spelling used in present-day dictionaries dominates among the corpus hits across the twentieth century. Particularly the compounds which are now preferably spelled solid in the dictionaries have developed exactly as expected: while only 56.9 per cent of these were spelled solid in 1931, the proportion rises to 86.3 per cent in 2006 with a parallel decrease both in

Table 5.78 *Spelling of to+day in chronologically ordered British English corpora*

	B-LOB (1931)	LOB (1961)	FLOB (1991)	BEo6 (2006)
O	4	5	1	6
H	330	68	9	0
S	14	300	266	270

hyphenations and open spellings. Possibly, this tendency to concatenate formerly hyphenated words, which must have arisen between 1932 and 1960, gave support to the idea that the spelling of English compounds develops in regular ways. Table 5.78 shows how the high-frequency word *to+day* was practically always hyphenated in 1931 but frequently spelled solid in 1961 and almost exclusively solid in corpora after that date. The data for *to+night* are very similar. Since this tendency affected a number of high-frequency words, it may have seemed to observers that it affected the vocabulary as a whole. While Hypothesis H₅ must be refuted for lack of statistical significance, patterns in the results for both tests suggest that a development from open via hyphenated to solid spelling is not completely uncommon.

5.8.3 Summary

Section 5.8 investigates diachronic variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Etymological origin of the constituents
- Heterogeneous etymological origin of the constituents
- Synchronically felt foreignness of the compound
- Age of the compound according to the OED
- Spelling variant used in the OED's earliest attestation
- Spelling development from the earliest attestation in the OED to the unanimous present-day dictionary spelling of the OHS₆₀₀ compounds.

Statistical testing revealed a strong effect of several independent variables on the dependent variable 'compound spelling' [OHS] for the OHS₆₀₀ compounds. The following hypotheses were confirmed:

- Compounds conveying the impression of being foreign words disfavour solid spelling. [H2]
- Old compounds (first attested by 1500) favour solid spelling. [H3]
- Recent compounds (first attested in 1900 or later) disfavour solid spelling. [H4]

In contrast to one of the most common ideas about English compound spelling (and in accordance with the actual expectations of the present study), the result for the following hypothesis contradicts the expectations:

- Compounds do not generally start with open spelling, then become hyphenated and then solid. [H5]

Yet another finding can be considered uncertain:

- Compounds containing one Germanic and one Romance constituent disfavour solid spelling. [H1] However, the difference to the next most common spelling is relatively small.

In other respects, the findings for Hypotheses H1 and H2 permit more precise results than expected:

- Compounds containing one Germanic and one Romance constituent favour open spelling. [H1]
- Compounds giving the impression of being foreign words favour open spelling and disfavour hyphenation. [H2]

In addition to the results for the hypotheses under consideration, several other findings emerged from the detailed analysis of the material:

- Compounds tend to conserve the spelling of their first attested use in the OED. [H5]
- First attestations in the OED are rarely spelled solid, whereas hyphenation and open spelling are almost equally likely. [H5]
- If compounds change their spelling, the tendency is usually towards a stronger rather than a weaker degree of concatenation. [H5]

5.9 Discourse Variables

In contrast to the previous sections, which discuss potential determinants of English compound spelling concerning the properties of compounds and their constituents mainly from a structural perspective, the following sections introduce linguistic variables beyond the word and

sentence levels, namely establishment in discourse, register and regional variety.

5.9.1 *Establishment in Discourse*

One potential factor influencing the spelling of English compounds is the surrounding discourse in the sense of the preceding and lexical items, which creates cohesion within a text (cf. Halliday and Hasan 1976). When a language user writes a text, one may therefore expect *priming* as the facilitation of “the processing of a stimulus (the ‘target’) . . . if a similar stimulus (the ‘prime’) has been processed previously” (Snider 2009: 815). The presence of a particular compound spelling variant early in a text will thus presumably raise the likelihood for the same spelling variant in the remainder of the text. Most probably, compound spelling consistency within a text is closely related both to the distance between the occurrences of the compounds in the text and to temporal constraints. A short distance with the corresponding higher activation levels (cf. Mondorf 2003: 285–286) should increase the probability that the speller will remember the precise variant used, subconsciously repeat the same spelling or at least be aware of having written the compound not so long ago. Since the investigation of the role of establishment in discourse with its many facets was not possible within the scope of the present research project, the clarification of this issue must be left to future research, ideally on unedited texts.

5.9.2 *Register*

The types and number of compounds in a text depend to a certain extent on its register, i.e. the “variety associated with a particular situation of use” (Biber and Conrad 2009: 6).²¹ Thus phrase compounds are particularly frequent in journalistic and specialised language, and conversations and

²¹ While open spelling does not conform to orthographic norms in German, it occurs relatively frequently in advertisements, product names, the names of publishers (e.g. *Gunter Narr Verlag*) and labels (Barz 1993: 167). This can be explained by aesthetic considerations governing typographic layout and design (Barz 1993: 169–170): since long chains of letters or hyphens may disturb the symmetry or the design of a package, the uncanonical open *‘Brillen Putztücher* ‘cleaning wipes for glasses’ may be found in prominent position on a package, while the small accompanying text contains the generally accepted solid spelling *Brillenputztücher*. Aesthetic reasoning seems to affect particularly the (non-)use of the hyphen: Rabinovitch (2007) mentions “designers’ distaste for its ungainly horizontal bulk between words” and observes that “people feel that hyphens mess up the look of a nice bit of typography”.

personal letters contain fewer compounds than more formal and abstract texts (Schmid 2011: 142), possibly because the former are closer to the oral modality than the latter (cf. Biber 1988: 56–57). Deliberate acts of creating compound spelling consistency (e.g. by consulting reference works) are time-consuming and therefore more likely to occur in texts which are important for the speller (i.e. in formal rather than informal registers), possibly targeted at a hierarchically higher reader and produced in a situation that permits dedicating time to spellchecking (e.g. when texts are copy-edited for publication). There is thus an interconnection between register and speed and the whole situational context, including the medium (cf. 5.11).

5.9.2.1 *Hypothesis I1 – Editing*

Edited texts follow the compound spelling tendencies codified in dictionaries more closely than unedited texts do. → Confirmed.

Edited texts are presumably more consistent in the spelling of English compounds than unedited texts (particularly if a fixed house style is followed), because copy-editors may dedicate a large amount of time and concentration to the systematic introduction or removal of hyphens and blanks. While the copy-editors' guidelines are likely to recommend the usage of one particular dictionary or to suggest uncontroversial spelling variants for individual compounds, unedited texts can be expected to be less systematic in this respect and to contain more non-canonical spellings. This is formulated as Hypothesis I1:

I1: Edited texts follow the compound spelling tendencies codified in dictionaries more closely than unedited texts do.

Hypothesis I1 was tested by comparing the spelling of the OHS_600 compound types in two corpora, namely BEO6 (as the most recent corpus of edited British English used) and the Blog Authorship Corpus (which contains unedited electronic language produced under no special constraints such as time or space, and which is not controlled for the variety of English; cf. 4.2). The percentages in Table 5.79 refer to the distribution of corpus results within each unanimous dictionary spelling variant, and the shaded cells mark the spelling variants used by all the dictionaries.

A chi-square test of the data carried out in *Excel* by means of the CHITEST function yielded a highly significant result ($p = 0.000$). As expected, the edited texts use the dictionary spellings in the clear majority of cases (e.g. 86.3 per cent solid corpus spellings of solid dictionary

Table 5.79 *Spelling of the OHS_600 compound types in corpora of edited versus unedited texts, combined with the usual dictionary spelling*

Dictionary	Corpus	Edited BEo6	Unedited Blog
O	O	87 (85.3%)	177 (62.5%)
	H	10 (9.8%)	59 (20.8%)
	S	5 (4.9%)	47 (16.6%)
H	O	57 (36.5%)	180 (40.6%)
	H	84 (53.8%)	178 (40.2%)
	S	15 (9.6%)	85 (19.2%)
S	O	11 (9.4%)	164 (38.7%)
	H	5 (4.3%)	73 (17.2%)
	S	101 (86.3%)	187 (44.1%)
TOTAL		375	1,150

compounds). While the unedited texts also favour the dictionary spellings (e.g. 44.1 per cent solid corpus spellings of solid dictionary compounds), the differences are less marked than in the edited texts, and hyphenated dictionary compounds are even slightly more likely to occur with open than with hyphenated spelling in the unedited texts. Hypothesis I1 can therefore be confirmed: edited texts follow the compound spelling tendencies codified in the dictionaries more closely than unedited texts do.

5.9.3 Regional Variety

That the regional variety of English has some influence on the spelling of English compounds is a common hypothesis in the literature (cf. Chapter 1).

5.9.3.1 Hypothesis I2 – British versus American English

British English uses more hyphens in compounds than American English. → Refuted.

Table 5.80 *Spelling of the OHS_600 compound types in British versus American English corpora*

	LOB (BE 1961)	Brown (AE 1961)	FLOB (BE 1991)	FROWN (AE 1991)
O	125 (37.8%)	128 (39.5%)	137 (38.3%)	135 (37.2%)
H	115 (34.7%)	94 (29.0%)	108 (30.2%)	110 (30.3%)
S	91 (27.5%)	102 (31.5%)	113 (31.6%)	118 (32.5%)
	331	324	358	363

As far as the two major varieties of English are concerned, American spelling (which mostly follows Webster’s 1806 and 1828 dictionaries) is claimed to be more standardised than British spelling (which was influenced by Dr Johnson’s 1755 dictionary; cf. Peters 2004: 511). With regard to compounds, Butcher (1992: 154) states that “American authors tend to use fewer hyphens than the British”, and Quirk et al. (1985: 1569) expand that “instead we find the items open or solid (more usually the latter) where BrE may use a hyphen” (e.g. in AmE *language retarded* vs. BrE *language-retarded*). This can be formulated as Hypothesis I2:

I2: British English uses more hyphens in compounds than American English.

This hypothesis was tested by conducting parallel analyses of FLOB vs. FROWN (1991) and LOB vs. Brown (1961; cf. 4.2). Table 5.80 summarises the number of OHS_600 compound types for which open, hyphenated and solid spelling was found in each corpus. The chi-square test performed in *Excel* by means of the function CHITEST finds no significant difference between the British English FLOB corpus and the American English FROWN corpus ($p = 0.948$), and the proportions of open, hyphenated and solid compounds in the 1991 corpora are almost identical in British English (with 30.2 per cent hyphenations) and American English (with 30.3 per cent hyphenations, i.e. even a few more). If we go back to 1961, the picture is slightly different: while British English uses more hyphenations (34.7 per cent) than American English (29.0 per cent), a chi-square test performed in *Excel* by means of the function CHITEST finds no significant compound spelling differences between the British English LOB

Table 5.81 *Spelling of the OHS_600 compound tokens in British versus American English corpora*

	LOB (BE 1961)	Brown (AE 1961)	FLOB (BE 1991)	FROWN (AE 1991)
O	2,263 (66.7%)	1,932 (62.1%)	1,962 (60.5%)	1,782 (58.3%)
H	330 (9.7%)	296 (9.5%)	319 (9.8%)	377 (12.3%)
S	799 (23.6%)	883 (28.4%)	961 (29.6%)	898 (29.4%)
	3,392	3,111	3,242	3,057

corpus and the American English Brown corpus ($p = 0.259$). As a consequence, Hypothesis I2 must be refuted: British English does not use more hyphenation in compounds than American English. This result contradicts the impression that may be gained by reading billboards, shop signs etc. in the United States, which seem to announce e.g. *homemade* or *home made* products far more frequently than the hyphenated spelling variant.

However, the result for the types is supported by an analogous analysis of the OHS_600 compound tokens (cf. Table 5.81). Even though a statistically significant difference can be observed between the 1961 corpora LOB and Brown (with $p = 0.000$ in the chi-square test), this cannot be attributed to hyphenation (which merely differs by 0.2 per cent) but to the slight preference for open spelling in LOB (BrE) and for solid spelling in Brown (AmE). In the 1991 corpora, the proportion of hyphenations is unexpectedly slightly higher in American English (12.3 per cent) than in British English (9.8 per cent). Since the results of the chi-square test for FLOB (BrE) vs. FROWN (AmE) are significant ($p = 0.006$), Hypothesis I2 can thus be refuted in all respects – at least for the established OHS_600 compounds.

In order to determine whether the less established compounds behave in a comparable manner, an additional study was carried out on the 525 biconstituent compounds which only occur in one of the present study's six dictionaries (383 with open, 82 with hyphenated and 60 with solid dictionary spelling), e.g. *milk+product*. Table 5.82 summarises the results for the types and Table 5.83 those for the tokens. Since the chi-square tests for 1961 do not reach statistical significance (with $p = 0.231$ for the

Table 5.82 *Spelling of the compound types in British versus American English corpora for the Master List compounds with only one occurrence in the dictionaries*

	LOB (BE 1961)	Brown (AE 1961)	FLOB (BE 1991)	FROWN (AE 1991)
O	44 (69.8%)	78 (80.4%)	52 (71.2%)	82 (71.3%)
H	15 (23.8%)	13 (13.4%)	15 (20.5%)	26 (22.6%)
S	4 (6.3%)	6 (6.2%)	6 (8.2%)	7 (6.1%)
	63	97	73	115

Table 5.83 *Spelling of the compound tokens in British versus American English corpora for the Master List compounds with only one occurrence in the dictionaries*

	LOB (BE 1961)	Brown (AE 1961)	FLOB (BE 1991)	FROWN (AE 1991)
O	362 (86.6%)	362 (89.2%)	319 (87.4%)	376 (85.5%)
H	38 (9.1%)	32 (7.9%)	23 (6.3%)	53 (12.0%)
S	18 (4.3%)	12 (3.0%)	23 (6.3%)	11 (2.5%)
	418	406	365	440

types and $p = 0.463$ for the tokens), the slightly higher proportion of hyphenation in British English compared to American English cannot be considered in the discussion of the results. The chi-square tests comparing the spellings in 1991 did not reach statistical significance for the types ($p = 0.827$) but yielded a significant result for the tokens ($p = 0.001$).

The analysis of both types and tokens for 1991 achieves converging results, which support the unexpected finding that hyphenation is more common in American English than in British English, regardless of how established the compounds are, and that Hypothesis I2 therefore needs to be refuted for present-day English.

5.9.4 Summary

Section 5.9 investigates discourse-related variables which were expected to exert some influence on the spelling of English compounds. The following variables were coded in the database:

- Compound frequency in corpora of edited vs. unedited texts
- Compound frequency in corpora of British English (LOB, FLOB) vs. American English (Brown, Frown)
- Compound frequency in corpora of electronic texts (blog corpus, chat corpus, text message corpus).

Statistical testing revealed an effect of one independent variable on the dependent variable ‘compound spelling’ [OHS] for the OHS_600 compounds. The following hypothesis was confirmed:

- Edited texts follow the compound spelling tendencies codified in the dictionaries more closely than unedited texts do. [I1]

The findings for the following hypothesis, by contrast, contradict the expectations:

- British English does not use more hyphens in compound spelling than American English. [I2] (At least not in the 1991 samples.)

5.10 Systemic Variables

According to de Saussure (1916/1959: 114–115), “the value of each term results solely from the simultaneous presence of the others” in a system and is thus determined by the term’s similarity and dissimilarity to others. The potential influence on the spelling of English compounds of all variables subsumed under the heading of *systemic variables* is related to the fact that they are part of the system of the English language: the existence of minimal pairs, emphasis and analogy.

5.10.1 Existence of Minimal Pairs

The spelling of English compounds is frequently discussed in connection with the avoidance of potential confusion in the form of minimal pairs (cf. 7.1). In this reading of *minimal pair*, the term refers not to “pairs of words in which a difference in meaning depends on the difference of one phoneme” (Roach 2000: 66), but rather to identical sequences of

graphemes in which a difference in meaning correlates with a mere difference in compound spelling variant; e.g. *airline* ('air transport company') vs. *air line* ('pipe supplying air'; cf. Waite 1995). However, the second member in such a minimal pair need not necessarily be a compound – it can also be a phrase, e.g. in the pairings *a light-blue hat* (which refers to a pale colour) as against *a light blue hat* (which refers to the hat's weight; cf. *GPO Style Manual* 2008: 79), or hyphenated *50-odd villagers* ('about 50') as against open *50 odd villagers* ('strange people'; cf. Merriam-Webster 2001: 115). However, since the systematic investigation of minimal pairs in all possible contexts of use of the target compounds would require extensive and highly error-prone speculation on potential opposition, it was not pursued in the present study.

5.10.2 *Emphasis*

Whenever a conventional spelling exists, this creates an unmarked situation, which facilitates processing (cf. 7.2) and permits the marking of atypical uses as a by-product. Exceptional emphasis can then be placed on a compound in a particular situation by choosing an orthographic variant which is atypical for that particular lexeme: thus solid or hyphenated spelling for usually open compounds will underline the idea that they represent a conceptual unity, whereas hyphenated or open spelling for otherwise solid compounds will draw attention to the meaning of the compounds' components, which should be recombined in a more literal meaning, e.g. in re-motivating puns (Käge 1980: 94): thus the verb *recover* usually means 'get better', but if it is used in the sense of 'cover once again', the spelling *re-cover* may actually be preferred (cf. Vallins 1954: 171). Unusual compound spelling (particularly concatenation) is also employed as a stylistic device in literary texts from the early twentieth century, e.g. by John Dos Passos, who frequently uses unusual solid spellings such as *railroadstation* (Ernst 2000: 106), or in James Joyce's novel *Ulysses*, which might have inspired Dos Passos (cf. Ernst 2000: 109–110). On a more general level, however, emphatic and re-motivating uses of compound spelling were expected to occur so infrequently that they were not investigated systematically in the present study.

5.10.3 *Analogy*

The most important system-related potential determinant of English compound spelling is analogy, whereby new spellings are based on existing

spellings of similar compounds within the system. For instance, the use of particular compound constituents may result in a preference for a particular spelling variant (cf. Sepp 2006: 123): thus novel compounds ending in *man* tend to use solid spelling, whereas compounds with *call* as the second constituent prefer open spelling by analogy to the existing set of compounds following that pattern (e.g. *house call* or *conference call*). This kind of word-specific information is also listed in style guides: the majority of the principles in Strumpf and Douglas' (1988: 56–58) review of hyphenation are linked to particular constituents, and the *GPO Style Manual* (2008: 76) advocates solid spelling for all compounds beginning with the nouns *book* and *school* or ending in *berry* and *plane*, among others. A crucial concept for the investigation of the role of analogy in English compound spelling is that of the *constituent family*, which Plag (2010: 244) defines as “the set of compounds that share the first, or the second, constituent with a given compound”. Novel compounds whose constituents have a large family are harder to reject as non-words (Van Jaarsveld, Coolen and Schreuder 1994), and the position of the shared constituent is also of importance (Krott, Baayen and Schreuder 2001: 90). Since the members of a constituent family may exhibit *constituent family bias* and favour e.g. a particular stress pattern (Plag 2010: 24), one may expect that compounds with a large left or right constituent family whose members mainly use open/hyphenated/solid spelling should also favour the dominant spelling. This hypothesis was tested in the form of two position-dependent hypotheses.

5.10.3.1 Hypothesis J1 – Left Constituent Family Size

Compounds favour the spelling variant with the highest type frequency in their left constituent family. → Confirmed.

The first expectation regarding the impact of the constituent family on spelling variant selection can be formulated as follows:

J1: Compounds favour the spelling variant with the highest type frequency in their left constituent family.

In order to test this hypothesis, the program CompSpell computed spelling-sensitive left constituent family size for all biconstituent compounds in the Master List (e.g. *week+end*) by retrieving all items in the headword list from the *Macmillan English Dictionary for Advanced Learners* (2007) beginning with the left constituent of the compound (*week*). A dictionary rather than a corpus was used so as to permit the search for open

compounds with varying combinations of part of speech. The MED was chosen because of its recency and the availability of the complete headword list. Neither the target compound itself nor the constituent on its own were considered in the calculation. A following letter raised the count for solid constituent family size, a following hyphen counted towards hyphenated constituent family size and a following space counted towards open constituent family size. Open compounds were considered constituent family members in order to treat all three spelling variants as equal (in spite of de Jong et al.'s 2002: 557 experimental result "that English open compounds do not belong to the morphological families of simplex words"). In order to avoid the incorrect retrieval of suffixations (*week+ly*), solid combinations of the left constituent followed by the grapheme sequences corresponding to the suffixes in Table A.7 (e.g. <ly>) were not considered – only if they were directly followed by additional letters (e.g. in the hypothetical compound ?*week+lyrics*). A spot search revealed that the disadvantage of missing solid compounds ending in *age* 'number of years', *ship* 'large boat' and *wise* 'sensible' (which are formally identical with suffixes in Table A.7) was clearly outweighed by the large number of incorrect hits avoided with this strategy. In order to test the predictive value of constituent family size, the results were summarised by coding the spelling variant with the largest left constituent family size [LS_r] as open (*o*), hyphenated (*h*), solid (*s*) or unclear (*u*; in the case of a draw between the two largest values). For example, the compound *youth+hostel* has four left constituent family members with open spelling (*youth club*, *youth culture*, *youth hostelling* and *youth worker*) but no hyphenated or solid left constituent family members. The spelling predicted by the left constituent family of *youth+hostel* is therefore open – a prediction which corresponds to the actual spelling of the compound in all the dictionaries under consideration.

In order to test Hypothesis J1, Pearson's chi-square test was carried out for the dependent variable 'compound spelling' [OHS] and the independent variable 'spelling variant with the highest type frequency in the left constituent family' [LS_r] on the OHS_600 compounds. The results are highly significant ($p = 0.000$). As expected, the spelling preferred by most left constituent family members tends to coincide with the spelling preferred by all the dictionaries (cf. the shaded cells in Table 5.84): 51.5 per cent of the predicted open spellings, 78.0 per cent of the predicted hyphenations and 48.2 per cent of the predicted solid spellings agree with the dictionary data. Hypothesis J1 can thus be confirmed: compounds favour the spelling variant with the highest type frequency in their left constituent family. The eighty-seven compounds whose orthography-sensitive left constituent

Table 5.84 *Spelling with the highest type frequency in the left constituent family [LS_r] and spelling in OHS_600*

		Spelling favoured by left constituent family size					
		open	hyphenated	solid	unclear	Total	
OHS	o	Count	101	2	52	45	200
		Expected Count	65.3	13.7	92.0	29.0	200.0
		% within LS_r	51.5%	4.9%	18.8%	51.7%	33.3%
	h	Count	54	32	91	23	200
		Expected Count	65.3	13.7	92.0	29.0	200.0
		% within LS_r	27.6%	78.0%	33.0%	26.4%	33.3%
	s	Count	41	7	133	19	200
		Expected Count	65.3	13.7	92.0	29.0	200.0
		% within LS_r	20.9%	17.1%	48.2%	21.8%	33.3%
Total	Count	196	41	276	87	600	
	Expected Count	196.0	41.0	276.0	87.0	600.0	
	% within LS_r	100.0%	100.0%	100.0%	100.0%	100.0%	

family size yields no clear spelling preferences are usually spelled open in the dictionaries. Note that the predictive accuracy differs for the three spelling variants: while open spellings are about as frequent as expected by chance (196 instead of 200), there is a clear bias towards solid spelling (276) and a clear avoidance of hyphenation (41) if spelling variant selection is based on spelling-sensitive left constituent family size alone.

5.10.3.2 Hypothesis J₂ – Right Constituent Family Size

Compounds favour the spelling variant with the highest type frequency in their right constituent family. → Confirmed.

By analogy to Hypothesis J₁, the expectation that constituent family size determines spelling variant selection can be applied to the second constituent of biconstituent compounds:

J₂: Compounds favour the spelling variant with the highest type frequency in their right constituent family.

Right constituent family size was determined by analogy to left constituent family size (cf. 5.10.3.1): CompSpell retrieved all MED headwords ending with the right constituents of all biconstituent Master List compounds (e.g. *end* for *week+end*). A preceding letter raised the count for solid constituent family size, a preceding hyphen counted towards hyphenated

Table 5.85 *Spelling with the highest type frequency in the right constituent family [RS_r] and spelling in OHS_600*

		Spelling favoured by right constituent family size					
		open	hyphenated	solid	unclear	Total	
OHS	o	Count	117	5	41	37	200
		Expected Count	62.7	29.7	77.0	30.7	200.0
		% within RS_r	62.2%	5.6%	17.7%	40.2%	33.3%
	h	Count	22	82	53	43	200
		Expected Count	62.7	29.7	77.0	30.7	200.0
		% within RS_r	11.7%	92.1%	22.9%	46.7%	33.3%
	s	Count	49	2	137	12	200
		Expected Count	62.7	29.7	77.0	30.7	200.0
		% within RS_r	26.1%	2.2%	59.3%	13.0%	33.3%
	Total	Count	188	89	231	92	600
		Expected Count	188.0	89.0	231.0	92.0	600.0
		% within RS_r	100.0%	100.0%	100.0%	100.0%	100.0%

constituent family size and a preceding space counted towards open constituent family size. As in the calculation of left constituent family size, apostrophes were treated like letters and thus counted towards solid spelling (e.g. in *M'Lord* as a solid right constituent family member of *land+lord*). In order to avoid the incorrect retrieval of prefixations (such as hypothetical [?]*mid+end* for *end*), solid combinations of the grapheme sequences corresponding to the prefixes in Table A.8 followed by the right constituent were not counted – only if they were directly preceded by additional letters (e.g. in the hypothetical compound [?]*pyramid+end*).

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'spelling variant with the highest type frequency in the right constituent family' [RS_r] yielded highly significant results ($p = 0.000$) for the OHS_600 compounds. As expected, the dominant spelling of the right constituent family members tends to coincide with the spelling of the OHS_600 compounds (cf. the shaded cells in Table 5.85): 62.2 per cent of the open, 92.1 per cent of the hyphenated and 59.3 per cent of the solid predicted spellings agree with the dictionaries. Hypothesis J2 is therefore supported by the data: compounds favour the spelling variant which has the highest type frequency in their right constituent family. The ninety-two compounds whose spelling-sensitive right constituent family size yields no clear spelling preferences are usually either

hyphenated (46.7 per cent) or spelled open (40.2 per cent) in the dictionaries. Note that the predictive accuracy differs for the three spelling variants: if spelling variant selection is only based on spelling-sensitive right constituent family size, open spellings are slightly less frequent than expected by chance (188 instead of 200), while there is a clear bias towards solid spelling (231) and a clear avoidance of hyphenation (89) – findings which parallel those for the left constituent family. If we compare how well position-dependent constituent family size predicts the spelling of the OHS_600 compounds, we find that right constituent family size outperforms left constituent family size with 336 as against 266 correct predictions.

5.10.3.3 *Hypothesis J3 – Left Constituent Family Frequency*

Compounds favour the spelling variant with the highest token frequency in their left constituent family. → Confirmed.

While the two previous hypotheses were based on the constituent family's size, it is also conceivable that the summed frequencies of the spelling-sensitive constituent family members should influence variant selection (cf. also Plag 2010). One may therefore expect that compounds whose constituent family members mainly occur with open/hyphenated/solid spelling (if individual usages are summed) should also favour open/hyphenated/solid spelling, respectively. This hypothesis was tested in the form of two position-dependent hypotheses, of which the first was formulated as follows:

J3: Compounds favour the spelling variant which has the highest token frequency in their left constituent family.

In order to test Hypothesis J3, spelling-sensitive left constituent family frequency was determined for each OHS_600 compound by an automated search in the written component of the British National Corpus. No lemmatised frequency list could be used, since only solid and hyphenated compounds yield lemmatised frequency data in BNCweb's CQP edition (bncweb.lancs.ac.uk/bncwebXML/Simple_query_language.pdf, 06 February 2013). The frequency data were extracted from a MySQL frequency list of BNCwritten two-grams by means of a Perl script. The joint frequencies for all open spellings of all constituent family members (but not the OHS_600 words themselves) were added together, and so were those for hyphenated and solid spellings, respectively. Like for the other analogical variables, the results were summarised by coding the spelling variant with the highest left

Table 5.86 *Spelling with the highest token frequency in the left constituent family [LF_r] and spelling in OHS_600*

			Spelling favoured by left constituent family frequency				Total
			open	hyphenated	solid	unclear	
OHS	o	Count	83	3	80	34	200
		Expected Count	54.7	10.0	114.3	21.0	200.0
		% within LF_r	50.6%	10.0%	23.3%	54.0%	33.3%
	h	Count	45	18	117	20	200
		Expected Count	54.7	10.0	114.3	21.0	200.0
		% within LF_r	27.4%	60.0%	34.1%	31.7%	33.3%
	s	Count	36	9	146	9	200
		Expected Count	54.7	10.0	114.3	21.0	200.0
		% within LF_r	22.0%	30.0%	42.6%	14.3%	33.3%
	Total	Count	164	30	343	63	600
		Expected Count	164.0	30.0	343.0	63.0	600.0
		% within LF_r	100.0%	100.0%	100.0%	100.0%	100.0%

constituent family frequency [LF_r] as open (*o*), hyphenated (*h*), solid (*s*) or unclear (*u*).

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'spelling variant with the highest token frequency in the left constituent family' [LF_r] yielded highly significant results ($p = 0.000$) for OHS_600. Fifty point six per cent of the open, 60.0 per cent of the hyphenated and 42.6 per cent of the solid spellings predicted by LF_r coincide with the actual spellings in the dictionaries (cf. the shaded cells in Table 5.86). Hypothesis J3 can therefore be confirmed: compounds favour the spelling variant which has the highest token frequency in their left constituent family. The compounds with no clear spelling bias based on left constituent family frequency are mainly spelled open (54 per cent). Note, however, that left constituent family frequency only predicts the spelling of 247 OHS_600 compounds correctly. This relatively low absolute number can be explained by a strong bias of left constituent family frequency in favour of solid spelling (343) and against hyphenation (30). This bias is partly due to the incorrect automatic coding of some MED headwords as solid constituent family members of short items, even though they are not compounds (e.g. **ear+lier* or **ear+nest* for the short left constituent *ear* in *ear+ache*). Since

constituent family frequency is based on constituent family size, it is presumably a less good predictor, because mistakes made in the determination of type frequency are carried over and complemented by additional mistakes in the computation of token frequency.

5.10.3.4 *Hypothesis J4 – Right Constituent Family Frequency*

Compounds favour the spelling variant with the highest token frequency in their right constituent family. → Confirmed.

By analogy to Hypothesis J3, the expectation that constituent family frequency determines spelling variant selection can be applied to the second constituent of biconstituent compounds:

J4: Compounds favour the spelling variant which has the highest token frequency in their right constituent family.

Right constituent family frequency was determined by analogy to left constituent family frequency (cf. 5.10.3.3).

Pearson's chi-square test for the dependent variable 'compound spelling' [OHS] and the independent variable 'spelling variant with the highest token frequency in the right constituent family' [RF_r] yielded highly significant results ($p = 0.000$) for OHS_600. Fifty-five point two per cent of the open, 89.4 per cent of the hyphenated and 49.7 per cent of the solid spellings predicted by RF_r (and thus the majority for each spelling variant) coincide with the actual spellings in the dictionaries (cf. the shaded cells in Table 5.87). Hypothesis J4 can therefore be confirmed: compounds favour the spelling variant with the highest token frequency in their right constituent family. The compounds with no clear spelling bias based on left constituent family frequency are mainly hyphenated (48.6 per cent). Altogether, 296 spellings are predicted correctly if the spelling variant with the highest token frequency in the right constituent family is chosen. Once again, there is a predictive bias in favour of solid spelling (for 306 out of 600 compounds) and against hyphenation (66). If we compare how well left and right constituent family frequency predict the spelling of the OHS_600 compounds, we find that right constituent family frequency outperforms left constituent family frequency with 296 as against 247 correct predictions.

Table 5.87 *Spelling with the highest token frequency in the right constituent family [RF_r] and spelling in OHS_600*

			Spelling favoured by right constituent family frequency				Total
			open	hyphenated	solid	unclear	
OHS	o	Count	85	5	79	31	200
		Expected Count	51.3	22.0	102.0	24.7	200.0
		% within RF_r	55.2%	7.6%	25.8%	41.9%	33.3%
	h	Count	30	59	75	36	200
		Expected Count	51.3	22.0	102.0	24.7	200.0
		% within RF_r	19.5%	89.4%	24.5%	48.6%	33.3%
	s	Count	39	2	152	7	200
		Expected Count	51.3	22.0	102.0	24.7	200.0
		% within RF_r	25.3%	3.0%	49.7%	9.5%	33.3%
Total	Count		154	66	306	74	600
	Expected Count		154.0	66.0	306.0	74.0	600.0
	% within RF_r		100.0%	100.0%	100.0%	100.0%	100.0%

5.10.4 *Summary*

Section 5.10 investigates systemic variables which were expected to exert some influence on the spelling of English biconstituent compounds. The following variables were coded in the database:

- Spelling-sensitive left constituent family size based on the MED
- Spelling-sensitive left constituent family frequency in BNCwritten
- Spelling-sensitive right constituent family size based on the MED
- Spelling-sensitive right constituent family frequency in BNCwritten.

Statistical testing revealed a strong effect of several independent variables on the dependent variable ‘compound spelling’ [OHS] in the OHS_600 sample. The following hypotheses were confirmed:

- Compounds favour the spelling variant with the highest type frequency in their left constituent family. [J1]
- Compounds favour the spelling variant with the highest type frequency in their right constituent family. [J2]
- Compounds favour the spelling variant with the highest token frequency in their left constituent family. [J3]
- Compounds favour the spelling variant with the highest token frequency in their right constituent family. [J4]

Table 5.88 *OHS_600 spellings predicted correctly by the constituent family variables*

		O	H	S	Total
RS_r	Right constituent family size	117	82	137	336
RF_r	Right constituent family frequency	85	59	152	296
LS_r	Left constituent family size	101	32	133	266
LF_r	Left constituent family frequency	83	18	146	247
CF	Combined left/right constituent family sizes/frequencies	70	27	137	234
CFS	Combined left and right constituent family sizes	62	13	87	162
CFF	Combined left and right constituent family frequencies	36	7	113	156

Since all analogical variables tested in the present study were found to have a significant effect on English compound spelling, their predictive accuracy was compared (cf. Table 5.88) in order to permit a ranking.

The combined constituent family sizes [CFS] and frequencies [CFF] clearly underperform the other variables. Merely the combination of all four constituent family variables [CF] achieves a result which is almost as good as that for the left constituent family values. Right constituent family size outperforms all other analogical variables with 336 correct predictions (= 56.0 per cent) and a relatively good performance for all three spelling variants. Since the next best result was achieved by right constituent family frequency (with 296 correct predictions), we can conclude that the right constituent plays an even more important role in the analogical spelling of English compounds than the left constituent. Not only do these results contradict Kuperman and Bertram (2013: 950), according to whom token-based constituent family frequency is “more predictive of the spelling alternation” than type-based constituent family size, but they are also counter-intuitive in view of the prominent role of the left constituent in lexical processing and the larger effect size of the left constituent family observed e.g. by Plag (2010: 276). However (and in contrast to Plag’s stress perspective, where the relative weight of main vs. secondary stress might necessarily require a more intense processing of the whole compound before its production), the potential first constituent of a compound is spelled identically regardless of whether it stands as a word of its own (with a following space or punctuation mark) or is followed by another constituent with which it enters into a compounding relation. The decision whether to concatenate, use a space or a hyphen may consequently depend more on the

second constituent as a potential run-on of the word (cf. Bell and Plag 2012: 510 for a similar argument regarding compound stress).

To sum up, if we compare the performance of all analogical variables, the following finding emerges:

- Orthography-sensitive right constituent family size is the best predictor of compound spelling among the analogical variables. [J1–J4]

However, constituent family data only predict hyphenation for 7 to 82 out of 200 compounds (depending on the variable under consideration), in spite of the fact that hyphenations should be as frequent as open and solid spellings. This may either mean that analogy is no useful guiding principle for compound spelling – or, what is more likely, that truly analogical influences require more than an identical first or last constituent; e.g. a similar underlying structure of the whole compound: while previous research has mainly considered noun+noun compounds, most of the hyphenations in OHS_600 begin with constituents other than a noun (145) or end with constituents other than a noun (151). An alternative explanation could therefore be that the variable ‘part of speech’ overrides the variable (or cluster of variables) ‘spelling preferred by the constituent family’ (cf. 7.3). This would mean that adjectives, for instance, would favour hyphenation by analogy to the general pattern found in other adjectives, in spite of any opposing analogical tendencies of their constituents.

5.11 Extralinguistic Variables

In contrast to the variables discussed so far, which are directly related to the compounds under consideration (e.g. part of speech) or at least to language (e.g. regional variety), the following sections discuss extralinguistic variables concerning the circumstances under which language is produced in the form of the language user, the medium and temporal and spatial restrictions.

5.11.1 *Language User*

While we have discussed compound spelling more or less independently from its users so far, spelling is actually a domain in which language users’ competence seems to vary considerably. With regard to English compound spelling, the following user-dependent variables are among those that can be expected to influence variant selection for particular subgroups of language users:

- **age:** older language users may continue to use the spelling they were taught at school, even if usage has changed. In view of the general tendency to use punctuation ever more sparsely (cf. e.g. Carey 1958: 5), which is reflected in the *Shorter Oxford English Dictionary's* (SOED, 2007) corpus-based respelling of ca. 16,000 former hyphenations as either open or solid compounds in its sixth edition (cf. SOED preface; Rabinovitch 2007), younger spellers might use fewer punctuation marks than older ones and consequently avoid hyphenation in compounds.
- **gender** and **class:** women and members of the supposed middle classes are claimed to have a stronger tendency to observe explicitly or implicitly recognised norms than men or members of the supposed upper and lower classes (cf. Coates 2004: 61–66). As a consequence, the former two groups are more likely to attempt to spell English compounds in what is considered the correct way.
- **native language:** the compounding conventions in the native language of speakers of English as a second language may influence their spelling variant selection. Thus native speakers of German may tend to overuse solid spelling, which is the standard for German compounds (Duden 2006: 1172–1183).

Many more user-related variables are conceivable, e.g. visual memory for specific spellings, memory for prescriptive orthographic rules, importance attached to consistency, personal preference or dispreference for one of the three orthographic variants – but also situation-related user variables such as limited attention span due to external phenomena (e.g. tiredness or distractions). Taking all of these variables systematically into account is, however, beyond the scope of the present research project, so that socio-linguistic user-dependent studies of English compound spelling are left to future research.

5.11.2 *Medium*

The media – sometimes even the material structure – in which compounds occur presumably also influence spelling variant selection. For instance, some unusual end-of-line hyphenations in Old English manuscripts can be explained by damages in the parchment or the wish to keep a straight margin (Wetzel 1981: 23–25), as the introduction of a hyphen makes a line longer and is exploited in terms of layout.

Presently, one may expect a large proportion of many language users' own written text production to occur in electronic media, which all have their specific characteristics, e.g. regarding how text is typed under temporal and spatial constraints. Since such considerations concern general principles of text production, medium was only considered indirectly by contrasting corpora containing text messages, blog posts and chat communication, respectively (cf. 5.11.3).

5.11.3 *Economy*

Given that language change is supposedly influenced by a principle of economy (cf. Brinton and Arnovick 2006: 56–57), one may expect that the sparing use of resources also plays some role in English compound spelling. This is particularly plausible considering that compounds are economical linguistic units, whose low production cost outweighs potential vagueness and ambiguity (Schmid 2011: 143). Economy may concern various aspects of compound spelling:

- using as little time as possible for the physical act of spelling (cf. 5.11.3.1)
- using as little effort as possible while spelling
- making the spelled compound maximally easy to process for readers
- making the spelled compound occupy as little physical space as possible (cf. 5.11.3.2)
- making the spelled compound use as little material as possible (which is closely related to using little space).

The effects of speed of typing and spatial restriction were tested in separate hypotheses.

5.11.3.1 *Hypothesis K1 – Speed of Typing*

Higher speed of typing favours open spelling. → Confirmed.

In the early twenty-first century, most written text is produced by digital means and read either in print or on a screen. The keyboards used for typing text into electronic devices such as computers or mobile phones, which influence speed of typing, may vary in structure (e.g. QWERTZ or QWERTY according to the default language) and they can be physical or virtual (e.g. on a smartphone, where the keyboards only appear when text is to be typed). Keyboards by different manufacturers share the basic arrangement of the letters of the alphabet. This goes back to the separation of common pairings on typewriters (as their predecessors) to minimise

jamming caused by adjacent keys, but while the arrangement could presumably be optimised for electronic keyboards, the traditional order has been conserved so far. Since the space bar at the bottom of the keyboard is usually larger than the other keys, the introduction of a space into a compound might almost be as fast as solid spelling, for which no additional key needs to be pressed. The hyphen, by contrast, is a small key of usual size and situated in the right little finger's area on both QWERTZ and QWERTY keyboards, which should make it more difficult and thus slower to type. Kuperman and Bertram (2013: 959) therefore assume that "hyphenated compounds would tend to become spaced when becoming more frequent, on the economy account". Open spelling, by contrast, may be the default option corresponding to the raw, unprocessed spelling of two (or more) words following each other, which is easy to type, common particularly in recent words (cf. 5.12.4), unmarked and not usually penalised by spellcheckers. All of these variables culminate in Hypothesis K1:

K1: Higher speed of typing favours open spelling.

In order to test this hypothesis, CompSpell automatically counted the number of open, hyphenated and solid hits for all Master List compound types in the Blog Authorship Corpus and the NPS Chat Corpus (cf. 4.2). While both corpora contain unedited informal electronic texts, a difference with regard to speed of typing was expected between blog posts (with no temporal restrictions for planning and typing) as against chat communication (typed in real time).

If we compare the spelling-sensitive hits for the OHS_600 compounds, the most striking result is the immense difference in the number of hits (cf. Table 5.89): 1,150 in the Blog Authorship Corpus as against 22 in the NPS Chat Corpus. This is due to the difference in size, with the blog corpus being about 3,637 times as large as the chat corpus. Nonetheless, the data permit the use of a chi-square test (performed in *Excel* by means of the function CHITEST). The results are statistically significant ($p = 0.009$), which can be explained by the marked tendency of the compound types in the chat corpus towards open spelling and against hyphenation. Hypothesis K1 can thus be confirmed: higher speed of typing favours open spelling.

If we compare the two corpora with regard to their compliance with the dictionary spellings (cf. Table 5.90), we find that the chat corpus's preference for open spelling leads to a spelling which differs from the dictionaries for all OHS_600 hyphenations. It is tempting to attribute this to the usual error induced by structurally identical phrases, but the manual search in

Table 5.89 *Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the NPS Chat Corpus*

	Low speed of typing (Blog)	High speed of typing (Chat)
O	521 (45.3%)	16 (72.7%)
H	310 (27.0%)	0 (0.0%)
S	319 (27.7%)	6 (27.3%)
TOTAL	1,150	22

Table 5.90 *Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the NPS Chat Corpus, combined with the usual dictionary spelling*

Dictionary	Corpus	Low speed of typing (Blog)	High speed of typing (Chat)
O	O	177	3
	H	59	0
	S	47	0
H	O	180	12
	H	178	0
	S	85	0
S	O	164	1
	H	73	0
	S	187	6

the corpus reveals that five of the twelve expected hyphenations with open spelling in the chat corpus are indeed compounds:

- ***Drive in?** I havent [sic] been to one of those since I was a kid*
- *did you attend that Kansas TC **get together?***
- *i just broke up with my gf.(**long term**).*
- ***snow white** :) [as a reply to *white lies*.]*
- *its a **win win** situation for me*

We may therefore conclude that there is a tendency of chat texts to use open spelling instead of hyphenation, whereas established solid compounds tend to be spelled solid.

Table 5.91 *Spelling of the OHS_600 compound tokens in the Blog Authorship Corpus and the NPS Chat Corpus*

	Low speed of typing (Blog)	High speed of typing (Chat)
O	115,528 (67.9%)	48 (85.7%)
H	6,496 (3.8%)	0 (0.0%)
S	48,201 (28.3%)	8 (14.3%)
TOTAL	170,225	56

If we consider the tokens, the difference between the blog and chat corpora is also statistically significant, with $p = 0.013$ in a chi-square test. While the absolute number of open spellings in both blog and chat texts is increased by comparison to the types (cf. Table 5.91), the proportion of open spellings remains higher in the chat corpus (85.7 per cent) than in the blog corpus (67.9 per cent).

This result is supported by a complementary study of the Master_5+ compounds, which provides additional evidence that open spelling is more frequent in the chat corpus (80.9 per cent) than in the blog corpus (41.7 per cent; $p = 0.000$ in a chi-square test). Hypothesis K1 can therefore be accepted: higher speed of typing favours open spelling. If spelling decisions have to be made under time pressure, recourse to a default (i.e. open spelling) is a very efficient solution – particularly in texts that are considered transitory. High speed of typing possibly correlates with a clear dispreference for hyphenation, because the use of a hyphen might slightly slow down the spelling process.

5.11.3.2 *Hypothesis K2 – Spatial Restriction*

Spatial restriction favours solid spelling. → Tentatively refuted.

The influence of spatial restriction on English compound spelling should be particularly obvious in text messages, a medium whose texts were originally limited to a length of 160 characters including punctuation. Particularly in the early days of texting (i.e. before the advent of flat rates and the opportunity to send longer text messages), it was financially advantageous for users to observe this limit. Since the awareness of writing in a spatially restricted medium may incite users to employ

Table 5.92 *Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the CorTxt Corpus*

	No spatial restriction (Blog)	Spatial restriction (SMS)
O	521 (45.3%)	43 (67.2%)
H	310 (27.0%)	1 (1.6%)
S	319 (27.7%)	20 (31.3%)
TOTAL	1,150	64

characters more economically than in other communicative contexts, the following hypothesis regarding the spelling of English compounds can be formulated:

K2: Spatial restriction favours solid spelling.

In order to test this hypothesis, CompSpell automatically counted the number of open, hyphenated and solid hits for all Master List compound types in the Blog Authorship Corpus and the CorTxt Corpus (cf. 4.2). While both corpora contain unedited informal electronic texts, a difference was expected between blog posts (with no spatial restrictions) as against text messages (with spatial restrictions and an expected increase in solid spelling). While the present-day predominance of flat rates permitting unlimited and extra-long text messages may have led to changes in texting behaviour, the collection period of the CorTxt Corpus from 2004 to 2007 (cf. 4.2) should ensure the original and prototypical spatial restrictions of the text message medium.

In contrast to the large number of hits for the OHS_600 compounds in the Blog Authorship Corpus (1,150 for all three spellings; cf. Table 5.90), only sixty-four compound types were found in the text message corpus – but this is not surprising in view of the fact that the former is about 140 times as large as the latter. A chi-square test carried out with the function CHITEST in *Excel* shows a significant difference between the blog and text message corpora regarding compound spelling ($p = 0.000$). This is mostly due to the quasi-absence of hyphenation from text messages (with the exception of *Well it was fifty-fifty!*), whereas almost one-third of the compound types retrieved from the blog corpus use a hyphen.

Table 5.93 *Spelling of the OHS_600 compound types in the Blog Authorship Corpus and the CorTxt Corpus, combined with the unanimous dictionary spelling*

Dictionary	Corpus	No spatial restriction (Blog)	Spatial restriction (SMS)
O	O	177	13
	H	59	0
	S	47	0
H	O	180	24
	H	178	1
	S	85	0
S	O	164	6
	H	73	0
	S	187	20

Unexpectedly, the proportion of solid spelling in text messages (31.3 per cent) corresponds almost exactly to the 33.3 per cent expected by chance, and the result is also close to the 27.7 per cent solid spellings in the blog corpus. Most of the compound types in text messages use open spelling (67.2 per cent), a proportion which is even higher than in the blog corpus (45.3 per cent). As a consequence, Hypothesis K2 has to be refuted: spatial restriction does not favour solid spelling.

While compound types with unanimous solid spelling in the dictionaries are predominantly spelled solid in the text message corpus (20 out of 26; cf. Table 5.93), this is only true of less than 50 per cent of the solid OHS_600 compound types in the blog corpus (187 out of 424). The general tendency of the few open and solid compounds in text messages to comply with the dictionary spelling is even stronger for the tokens (cf. Table 5.94).

For the tokens (as for the types), the difference in compound spelling between the blog corpus and the text message corpus was significant ($p = 0.000$ in a chi-square test carried out with the function CHITEST in *Excel*; cf. Table 5.95). At first sight, the results for the tokens seem to confirm Hypothesis K2, as a majority of solid spellings (55.1 per cent) in the text message corpus contrasts with a majority of open spellings in the blog corpus (67.9 per cent). However, the results for the text messages are due to a small number of recurrent compounds: there are only twenty types with solid spelling (cf. Table 5.92), of which *weekend* on its own represents 340 of the 427 solid tokens and thus skews the results. At the same time, only

Table 5.94 *Spelling of the OHS_600 compound tokens in the Blog Authorship Corpus and the CorTxt Corpus, combined with the unanimous dictionary spelling*

Dictionary	Corpus	No spatial restriction (Blog)	Spatial restriction (SMS)
O	O	8,442	22
	H	270	0
	S	475	0
H	O	104,195	314
	H	5,949	1
	S	2,443	0
S	O	2,891	11
	H	277	0
	S	45,283	427

Table 5.95 *Spelling of the OHS_600 compound tokens in the Blog Authorship Corpus and the CorTxt Corpus*

	No spatial restriction (Blog)	Spatial restriction (SMS)
O	115,528 (67.9%)	347 (44.8%)
H	6,496 (3.8%)	1 (0.1%)
S	48,201 (28.3%)	427 (55.1%)
TOTAL	170,225	775

0.1 per cent of the OHS_600 compound tokens in text messages are hyphenated, whereas open spelling is relatively frequent (44.8 per cent).

All of these findings can be linked to a joint explanation, namely the text recognition software used in mobile phones. Current versions are relatively sophisticated and automatically add words typed by the user to their database, but one may expect that the contributors to the CorTxt corpus used earlier versions of such software. Based on the characters typed in so far, text recognition software suggests complete lexemes for quick and convenient selection. Since many of the suggestions are simplex words rather than full compounds, this may incite writers of text messages who are spelling a compound to select and accept a suggested

first constituent and to move on to the suggestions for the next constituent. In this case, a space is inserted either manually or automatically, thereby resulting in open spelling of the compound. Impressionistically gathered evidence suggests that many users simply accept what the mobile offers them immediately, even if it is not precisely what they are looking for – which might explain the large proportion of open spellings. At least in some present-day mobiles, it is possible to insert a hyphen and receive suggestions for the next constituent, but hyphens are still more difficult to type than spaces, since they might require the additional selection of a punctuation mark option or the pressing of a smaller key on a smartphone. As a consequence, the extreme underuse of hyphenation in text message compounds can be expected to continue. By contrast, current text recognition software may suggest established solid compounds in solid spelling (e.g. *weekend*) and thus offer the most appropriate choice.

Taking all of this into account, Hypothesis K2 needs to be refuted: spatial restriction does not favour solid spelling. This result is in line with Bieswanger (2007), who finds an average text message length of only ninety-one characters in his corpus and concludes that space is no limiting factor in English text messages. Shortenings like LOL ('laughing out loud'; cf. LDOCE) are used less to save space than for reasons of playfulness and in order to appear witty (Bieswanger 2007). Since people in Western societies (particularly those with English as a native language) tend to read more now than a century ago, and since electronic texts (in which saving space is less important than in print) constitute a very large proportion of everyday written language use, economy of production (cf. Kuperman and Bertram 2013: 941) may have been replaced with economy of processing. If fast reading is the new guiding principle determining variant selection, open compound spelling might prove particularly suitable, as Inhoff, et al.'s (2000: 45) results for German suggest.

5.11.4 Summary

Section 5.11 investigates extralinguistic variables which were expected to exert some influence on the spelling of English compounds. The following variables were coded in the database:

- Temporal restrictions: compound frequency in blog posts (unmarked) vs. chat communication (temporally limited)

- Spatial restrictions: compound frequency in blog posts (unmarked) vs. text messages (spatially limited).

Statistical testing revealed an effect of one independent economical variable on the dependent variable ‘compound spelling’ [OHS] in the OHS_600 sample. The following hypothesis was confirmed:

- Higher speed of typing favours open spelling. [K1]

By contrast, the findings for the following hypothesis partly contradict the expectations:

- Spatial restriction does not favour solid spelling. [K2]

In addition to the results for the hypotheses under consideration, the following findings emerged from the detailed analysis of the material:

- Higher speed of typing disfavours hyphenation. [K1]
- Spatial restriction disfavours hyphenation. [K2]

5.12 General Issues

The previous sections tested the influence of a large number of variables on English compound spelling. These individual variables can be summarised in the form of more general super-variables:

- Heterogeneity of the constituents:
 - a large difference in the number of letters
 - a large difference in the number of syllables
 - a large difference in the frequencies
 - the combination of Germanic and Romance origin
 - the combination of synchronically felt foreignness with its absence
 - the combination of lexical and grammatical parts of speech
- Complexity of the compound:
 - more than two constituents
 - large number of letters
 - large number of syllables
 - presence of one or more complex constituents (prefixation, suffixation, acronym)

- Lexicalisation of the compound:
 - high frequency
 - old age (i.e. early first attestation)
 - idiomaticity
- Reduced readability at the constituent joint:
 - identical letters on both sides
 - consonant cluster across the constituent joint
 - garden path cluster across the constituent joint
 - vowel graphemes on both sides.

The following sections determine to what extent the spelling of English compounds is determined by these super-variables, which are frequently mentioned in the literature.

5.12.1 Hypothesis L1 – No Chaos

The present-day spelling of English compounds is influenced by a number of variables. → Confirmed.

The most central question which the present study attempts to answer is whether the spelling of English compounds is really arbitrary (cf. Chapter 1) or whether there are any underlying principles. The expectation that there is some regularity can be formulated in the form of the following hypothesis:

L1: The present-day spelling of English compounds is influenced by a number of variables.

Hypothesis L1 underlies the testing of practically all of the individual hypotheses presented earlier, so that any significant correlation between the dependent variable ‘English compound spelling’ and any independent variable (such as compound length) contributes to the confirmation of Hypothesis L1. Since the previous sections found a significant effect of numerous variables favouring or disfavouring one or more of the three spelling variants (cf. Table A.9 in the Appendix for an overview of these variables and their effect), Hypothesis L1 can be accepted: the spelling of English compounds is not completely unsystematic but influenced by a large number of variables from all of domains under consideration (spelling, length, frequency, phonology, morphology, grammar, semantics, diachronic variables, discourse variables, systemic variables and extralinguistic variables).

5.12.2 Hypothesis L2 – Heterogeneous Constituents

Heterogeneous constituents disfavour solid spelling. → Confirmed.

Several of the hypotheses tested earlier involve the expectation that spellers are reluctant to concatenate very dissimilar constituents. This can be formulated as the more general Hypothesis L2:

L2: Heterogeneous constituents disfavour solid spelling.

This hypothesis was tested by formalising the dissimilarity of a compound's constituents as involving any of the following:

- a) a ratio exceeding 2:1/1:2 in the number of letters (cf. 5.2.5)
- b) a ratio exceeding 2:1/1:2 in the number of syllables (cf. 5.2.6)
- c) a ratio exceeding 50:1/1:50 in the frequency values (cf. 5.3.4)
- d) a combination of a lexical constituent with a grammatical constituent (cf. 5.6.1.8)
- e) a combination of one constituent with Germanic origin and one with Romance origin (cf. 5.8.1.1).

The consideration of the results for each of the variable-specific sub-hypotheses a) to e) for OHS_600 shows that Hypothesis L2 can be accepted:

- a) A ratio exceeding 2:1/1:2 in the number of letters clearly **disfavours solid spelling** (1.8 per cent), clearly favours hyphenation (76.4 per cent) and slightly disfavors open spelling (21.8 per cent).
- b) A ratio exceeding 2:1/1:2 in the number of syllables clearly **disfavours solid spelling** (0.0 per cent), clearly favours open spelling (67.7 per cent) and does not affect hyphenation (32.3 per cent).
- c) A ratio exceeding 50:1/1:50 in the frequency values **disfavours solid spelling** (with values between 5.9 per cent and 20.6 per cent for various frequency ranges), favours hyphenation (with values between 47.6 per cent and 88.2 per cent) and either disfavors or has no effect on open spelling (with values between 5.9 per cent and 33.3 per cent).
- d) A combination of a lexical constituent with a grammatical constituent clearly **disfavours solid spelling** (9.1 per cent) and open spelling (6.1 per cent), while clearly favouring hyphenation (84.8 per cent).
- e) A combination of one Germanic and one Romance constituent **very slightly disfavors solid spelling** (27.3 per cent) and hyphenation (31.1 per cent), while favouring open spelling (41.5 per cent).

Hypothesis L2 is thus confirmed by the data: on all levels under consideration, heterogeneous constituents disfavour solid spelling. This finding receives further support from the observed but statistically non-significant tendency not to concatenate compounds if they contain some element that is not a regular lower-case letter, such as a numeral or a capitalised acronym (cf. 5.1.3 and 5.5.2.1). Since two of the five sub-hypotheses outlined earlier favour open spelling and three favour hyphenation, these two spelling variants seem to fulfil the task of maintaining very different constituents apart about equally well.

5.12.3 Hypothesis L3 – Complex Compounds

Complex compounds disfavour solid spelling. → Confirmed.

The next general hypothesis is based on the assumption that a large degree of complexity makes the recognition of a compound's constituents more difficult. Complexity may relate to the presence of a word formation (e.g. prefixation, suffixation) among the constituent(s) or to the compound's length (measured in constituents, syllables or letters): since complex compounds tend to be longer, their reading tends to require more fixations on the compound, and the readers' structural analysis may be simplified by providing visually salient spaces or hyphens between the constituents (cf. 1.1.1). One may therefore formulate the following hypothesis:

L3: Complex compounds disfavour solid spelling.

The consideration of the variables and their values which represent complexity yields the following results for the OHS_600 compounds:

- a) Compounds comprising more than two constituents **disfavour solid spelling** (those with three constituents favour open spelling; those with four constituents slightly favour hyphenation; cf. 5.2.1).
- b) Compounds comprising more than two syllables **disfavour solid spelling**: the trisyllabic ones prefer open spelling (45.6 per cent) over hyphenation (37.3 per cent), whereas longer compounds clearly favour hyphenation (73.6 per cent) over open spelling (26.4 per cent) and never use solid spelling (cf. 5.2.2).
- c) Compounds comprising more than ten letters clearly **disfavour solid spelling** (4.8 per cent) and clearly favour open spelling (60.3 per cent; cf. 5.2.3).
- d) Compounds containing one or more complex constituents (prefixations, suffixations or acronyms) clearly **disfavour solid spelling** (2.6

per cent) and favour hyphenation (58.3 per cent) over open spelling as the second most frequent variant (39.1 per cent; cf. 5.5.2.1).

Hypothesis L3 is thus confirmed by the data: on all levels under consideration, complex compounds disfavour solid spelling. The choice between open spelling and hyphenation is not unanimous, but all of the foregoing variables considered jointly seems to imply that compounds with a particularly high degree of complexity tend to be hyphenated – possibly because this permits easy segmentation while conserving orthographic unity.

5.12.4 Hypothesis L4 – Lexicalised Compounds

Lexicalised compounds favour solid spelling. → Tentatively refuted.

A complex variable frequently discussed in the context of compound spelling is that of *lexicalisation* – e.g. when Mondorf (2009: 375) states that “the degree of lexicalisation is mirrored in compound spelling”; following Lipka’s (1977: 55) definition as “the process whereby frequently used morphologically complex lexemes tend to become single lexical entities with the concomitant effects of gaining a more specific meaning and losing some of their structural transparency” (Mondorf 2000: 36). Other uses of the term (e.g. Bauer 1983: 48–49) understand *lexicalisation* as the final stage in the history of a lexeme, in which a word formation has become opaque with regard to its phonology, morphology or semantics, or in which the rules that have led to its formation are no longer productive. Since the loss of transparency is inherently gradual (cf. also Sanchez 2008), the term *established* is sometimes used, which subsumes *lexicalisation* and its preceding stage of *institutionalisation* (in which a lexical item is recognised by other speakers as known and thus item-familiar; cf. Bauer 1983: 48–50). Since the idea that English compounds develop from an open via a hyphenated stage to solid spelling is relatively widespread (cf. 5.8.2.3), we can formulate the following hypothesis:

L4: Lexicalised compounds favour solid spelling.

Lexicalisation can be operationalised as comprising any of the following three features: high frequency, an early date of first attestation and/or a certain degree of idiomaticity of the compound. The older and the more frequent a compound, the likelier it is to be an integral part of the lexicon, and the higher the probability that it will be spelled as a single word.

The consideration of the variables and their values which represent lexicalisation yields the following results for the OHS_600 compounds:

- a) Very high compound frequency is likely to result in solid spelling (40.1 per cent). However, the proportion of hyphenations is even higher for such compounds (40.8 per cent; cf. 5.3.1).
- b) Idiomatic compounds (*i*) favour solid spelling (40.7 per cent), but the proportion of hyphenations is close (38.3 per cent). Highly idiomatic compounds (*b*), by contrast, clearly favour hyphenation (43.1 per cent) and disfavour open spelling (25.0 per cent). In this group, the proportion of solid spelling (31.9 per cent) is very close to the expected average (cf. 5.7.3.1).
- c) Compounds with an early first attestation (1500 or earlier) favour solid spelling (67.8 per cent), while disfavouring both hyphenation (20.3 per cent) and particularly open spelling (11.9 per cent; cf. 5.8.2.1).

Based on these results, it is difficult to clearly accept or refute the hypothesis that lexicalised compounds favour solid spelling: while Hypothesis L₄ is supported by the results regarding the age of the compound and an intermediate degree of idiomaticity, the hypothesis is unexpectedly contradicted by the data for highly idiomatic compounds, and the results are slightly inconclusive for high-frequency compounds (where solid spelling comes second by an extremely small margin). Taking all these findings into account, it was decided to refute Hypothesis L₄: lexicalised compounds do not favour solid spelling. However, one may argue that the sample under consideration is not ideal for the testing of this particular hypothesis, since all of the OHS_600 compounds are lexicalised by definition if inclusion in many dictionaries is equated with lexicalisation. Nevertheless, this can be remedied to a certain extent by using the frequency of occurrence in different dictionaries as an indicator of the degree of lexicalisation.

Table 5.96 *Lexicalisation and spelling*

Sample	Degree of lexicalisation	Spelling variation	O	H	S	Total types
OHS_600	Relatively high	–	200 (33.3%)	200 (33.3%)	200 (33.3%)	600
OHS_extra	Relatively high	–	2,317 (60.0%)	325 (8.4%)	1,222 (31.6%)	3,864
Master_5+	Relatively high	+	1,858 (36.8%)	1,560 (30.9%)	1,631 (32.3%)	1,196
Master_1–4	Relatively low	+	7,922 (69.7%)	1,754 (15.4%)	1,685 (14.8%)	3,903

We can observe a similar proportion (about one-third) of solid spellings in all compound lists whose items occur in at least five dictionaries (OHS_600, OHS_extra, Master_5+), regardless of whether the items are spelled identically in all the dictionaries. Since solid spelling is clearly less frequent in the Master_1–4 compounds with a supposed low degree of lexicalisation (14.8 per cent), these results suggest that solid spelling is more commonly found in lexicalised compounds indeed. However, even for the compounds with the highest degree of lexicalisation, the proportion of solid spelling is only about one-third, whereas open spelling is so common (with 60 per cent for OHS_extra) that these results provide additional justification for the rebuttal of Hypothesis L4.

5.12.5 *Hypothesis L5 – Obstacles to Readability*

Reduced readability across the constituent joint disfavours solid spelling. → Tentatively refuted.

Style guides frequently offer the advice not to use solid spelling if it creates visual obstacles for the reader (cf. 5.1.1). This can be reformulated in the form of the following general hypothesis:

L5: Reduced readability across the constituent joint disfavours solid spelling.

In the empirical study, various types of obstacle were tested with regard to their influence on compound spelling, with the following results:

- a) Identical letters across the constituent joint clearly **disfavour solid spelling** (5.3 per cent = one hit) and favour hyphenation (52.6 per cent = ten hits) or open spelling (42.1 per cent = eight hits). Since the numbers are so small, the difference between the last two results may be a matter of chance.
- b) Clusters of four or more consonants across the constituent joint unexpectedly **do not disfavour solid spelling**: with 40.7 per cent (= twenty-four hits) each, solid spelling and hyphenation achieve identical top results, whereas open spelling is less common (18.6 per cent = eleven hits).
- c) Garden path clusters across the constituent joint are so rare in the OHS_600 data that the OHS_extra compounds had to be considered. Unexpectedly, **solid spelling was more frequent than expected** (24 instead of 16.1 expected counts), whereas hyphenation was close to

expectations and open spelling less frequent than expected (22 instead of 30.6 expected counts).

- d) Vowel graphemes on both sides of the constituent joint clearly **disfavour solid spelling** (3.8 per cent). Open spelling remains largely uninfluenced (34.6 per cent), but hyphenation is clearly increased (61.5 per cent).

Since the results are inconclusive, Hypothesis L5 is tentatively refuted: reduced readability across the constituent joint does not always disfavour solid spelling. However, since all variables referring to length-related complexity do disfavour solid spelling (cf. 5.12.3), reduced readability for the complete compound may be assumed to play a role instead.

5.12.6 Hypothesis L6 – Direction of Tendency

The variables have a stronger tendency to favour than to disfavour one of the three spelling types. → Confirmed.

If one considers the aforementioned four general hypotheses, all of which make predictions regarding the spelling of specific groups of compounds, it is striking that three predict the avoidance of solid spelling (L2, L3 and L5), whereas only one expects maximal orthographic unity (L4). This raises the question whether the results for individual variables tend to favour or disfavour one particular spelling. In the latter case, this would result in more uncertainty regarding the most appropriate variant. Hypothesis L6 formulates the following expectation:

L6: The variables have a stronger tendency to favour than to disfavour one of the three spelling types.

Hypothesis L6 was tested by analysing Table A.9 in the Appendix, which summarises how the seventy-five tested variables/values with statistically valid results correlate with compound spelling. It was found that sixty-four cases favour one spelling type (with a very marked difference in thirty instances) and eleven cases favour two of the three variants to a highly similar degree. Hypothesis L6 is thus confirmed by the data: the variables have a stronger tendency to favour than to disfavour one of the three spelling variants. If we consider the favoured spelling variants in detail, we find the following tendencies (cf. Table 5.97): at least for those combinations of variables and values which were of interest for the overview, we can state that open spelling is relatively unbiased (with thirty preferences compared to thirty-eight dispreferences). Hyphenation

Table 5.97 *General spelling tendencies extracted from Table A.9*

Effect		O	H	S
++	strong preference	11	21	3
+	preference	19	18	19
o	no or little effect	7	12	7
-	avoidance	21	22	21
--	clear avoidance	17	2	25

is favoured more often than it is disfavoured (with thirty-nine as against twenty-four instances), whereas solid spelling is more likely to be avoided than to be preferred (with forty-six as against twenty-two instances). The results for solid spelling (which is mainly dispreferred) are particularly interesting, because they might imply that contrary to expectations, compounds with particular features are unlikely to progress into the direction of concatenation. Should the system not change, this would mean that we are not experiencing a period of transition, but that some compounds may never reach the generally assumed final state of solid spelling. This result supports Peters' (2004: 119) statement that some compounds, "especially longer ones like *daylight-saving*, may never progress beyond the hyphenated stage (in British English, or spaced, in American), however well established they are" – an aspect that has largely been overlooked in the literature so far (cf. also 5.8.2.3).

5.12.7 *Summary*

Section 5.12 investigates general hypotheses which are related to the spelling of English compounds by recombining the results of the previous sections. The following hypotheses were confirmed:

- The spelling of English compounds is not completely unsystematic but influenced by several variables. [L1]
- Heterogeneous constituents disfavour solid spelling. [L2]
- Complex compounds disfavour solid spelling. [L3]
- The variables have a stronger tendency to favour than to disfavour one of the three spelling types. [L6]

The results for the following hypotheses, by contrast, contradict the expectations:

- Lexicalised compounds do not favour solid spelling. [L4]
- Reduced readability across the constituent joint does not disfavour solid spelling. [L5]

If we reverse the perspective of Hypotheses L2 to L5 (which assume that compounds with particular qualities tend to be spelled in a particular way), we can ask whether it is possible to attribute any meaning to the three compound spelling variants by considering what types of variable/value combinations favour open, hyphenated and solid spelling, respectively, and by finding linking concepts beyond those formulated in Hypotheses L2 to L5. Considering the results for these hypotheses, compounds with open spelling are characterised by the following features (among others):

- relatively long compound
- relatively long constituents
- low compound frequency
- head-final morphological structure
- heterogeneous constituents
- literal meaning
- relatively recent date of first attestation.

Prototypical exemplars in this sense are, for example, *calendar month*, *carriage clock*, *elastic band* and *precious stone*. Interestingly, all of the foregoing features are reminiscent of phrases – even for the highly established compounds occurring in many different dictionaries. This seems to imply that open spelling does indeed convey a phrase-like meaning.

Hyphenations, by contrast, are characterised by the following features (among others):

- vowels across the constituent joint
- short constituents
- very large length difference between constituents (in number of letters)
- low constituent frequency
- very large frequency difference between constituents
- back stress
- adjective compound
- one or more complex constituents
- no head-final morphological structure
- high level of idiomaticity.

It was not possible to find a single compound comprising all these features in OHS_600. That is due, among other things, to the fact that a low

constituent frequency and a very large frequency difference between the constituents contradict each other. Similarly, complex constituents are usually not short. The category of hyphenated compounds therefore seems to be a mixed bag, in which merely some features apply to individual compounds in the sense of a family resemblance (cf. Wittgenstein 1972: 31–32, according to whom the members of the category ‘games’ do not all share one particular feature but are linked instead by “a complicated network of similarities overlapping and criss-crossing”). Nonetheless, on a more general level, it is also possible to recognise firstly that many hyphenated compounds are marked as somehow unusual (e.g. in terms of their morphological structure) and secondly that the majority of the hyphenations are adjectives. Compounds such as *red-faced*, *part-time* or *two-bit* are therefore relatively prototypical. Since hyphens correlate with a number of characteristics and are not generally used at random in free variation, they cannot be dispensed with as easily as suggested by Burridge (2005: 163).

Last but not least, solid compounds are characterised by the following features (among others):

- very short compound
- short constituents
- fore-stress
- simple constituents
- early date of first attestation.

All this supports traditional ideas about the category of solid compounds, for which *handbook*, *godson* and *workday* are typical exemplars.

5.13 Summary

Previous studies on the spelling of English compounds have considered a small number of variables for a large number of words. The present study, by contrast, analyses an extremely large number of potential variables in order to determine which ones play a role in determining the spelling of English compounds. Since many of these variables, particularly the more complex ones (such as morphological structure or etymological origin), had to be coded manually, the coding of the complete set of variables was restricted to the 600 randomly selected OHS_600 compounds, of which 200 each occur with exclusively open, hyphenated and solid spelling in at least five dictionaries. In addition, those variables that were coded either automatically or semi-automatically were coded for all compounds from the complete Master List. Table 5.98 gives an overview of all the variables

Table 5.98 Variables coded in the database

VARIABLE	METHOD	SAMPLE
A) Spelling		
Consonant clusters: number of consonant graphemes across the constituent joint	CompSpell count	Master List
Identical graphemes before and after the constituent joint (<i>felt+tip</i>)	CompSpell search	Master List
Garden path clusters: misleading digraph across the constituent joint (<th> in <i>ant+hill</i>)	CompSpell search	Master List
Vowel graphemes before and after the constituent joint (<i>amino+acid</i>)	CompSpell search	Master List
Capitalisation of one or more constituents (<i>all+American</i>)	Coded manually following a case-sensitive search of the letters of the alphabet	Master List
Occurrence of one or more apostrophes within the compound (<i>seller's+market</i>)	Coded manually following an automatic search for apostrophes	Master List
B) Length		
Number of constituents of the compound	Coded manually following automatic sorting	Master List
Number of letters of the compound	CompSpell count	Master List
Number of letters of the constituents	CompSpell count	Master List
Ratio between the lengths of the constituents (letters)	<i>Excel</i> formula: a/b	Master List
Very different length of the first two constituents (letters)	Coded manually following automatic sorting	Master List
Number of syllables of the compound	CompSpell count	Master List
Number of syllables of the constituents	CompSpell count	Master List
Ratio between the lengths of the constituents (syllables)	<i>Excel</i> formula: a/b	Master List
Very different length of the first two constituents (syllables)	Coded manually following automatic sorting	Master List
C) Frequency		
Frequency of the compound with corresponding part of speech in the six dictionary lemma lists (LDOCE, MED etc.)	CompSpell search in the dictionary files	Master List
Frequency of the compound in corpora (Brown, LOB etc.)	CompSpell search in the corpora	Master List
Frequency of the compound in BNCwritten	Copying of external automated search results into the database	Master List

Table 5.98 (*cont.*)

VARIABLE	METHOD	SAMPLE
Frequencies of the compound constituents in the lemmatised BNCwritten frequency list	CompSpell search in the lemmatised BNCwritten frequency list <ul style="list-style-type: none"> • for OHS_600 with part of speech • for non-OHS_600 <ul style="list-style-type: none"> ◦ with the compound's part of speech for the last constituent ◦ without part of speech for all other constituents 	Master List
Ratio between the frequencies of the constituents in the lemmatised BNCwritten frequency list	<i>Excel</i> formula: a/b	Master List
Very different frequency of the first two constituents	Coded manually following automatic sorting	Master List
D) Phonology		
Constituent-final silent <e> in non-compound-final constituents	Coded semi-automatically	OHS_600
Stress pattern (main stress on first or second constituent)	<ul style="list-style-type: none"> • Coded by copying the stress patterns from the MED list covering 5,273 of the Master List compounds • Remaining OHS_600 compounds coded manually based on MED and Longman Pronunciation Dictionary 	Master List (part) OHS_600
E) Morphology		
Prefix in non-initial constituent	Coded manually	OHS_600
-ing, -ed or -er in compound-final position	<ul style="list-style-type: none"> • Regular forms coded manually for complete Master List • Irregular forms coded manually for OHS_600 	Master List OHS_600
Lexical suffix in non-final constituent	Coded manually	OHS_600
Inflection in non-final constituent	Coded manually	OHS_600

Table 5.98 (*cont.*)

VARIABLE	METHOD	SAMPLE
Complex constituents	<ul style="list-style-type: none"> • Capitalised acronyms coded semi-automatically for complete Master List • Other complex constituents coded manually for OHS_600 	Master List OHS_600
Morphological structure (head-initial, head-final, non-headed)	Coded manually	OHS_600
F) Grammar		
Part of speech of the compound	Standardisation of codes from dictionary headword files	Master List
Part of speech of the constituents	Coded manually	OHS_600
Part-of-speech class of the compound (open/lexical vs. closed/grammatical)	Coded semi-automatically	Master List
Part-of-speech class of the constituents (open/lexical vs. closed/grammatical)	Coded semi-automatically	OHS_600
Frequency in attributive, predicative and non-attributive position in BNCwritten	Copying of external automated search results into the database	Master List
G) Semantics		
General nouns as constituents	Coded semi-automatically	Master List
Semantic relation between the constituents	Coded manually	OHS_600
Idiomatcity of the compound	Coded manually	OHS_600
H) Diachronic variables		
Etymological origin of the constituents	Coded manually based on the OED	OHS_600
Heterogeneous etymological origin of the constituents	Coded semi-automatically based on the etymological origin of the constituents	OHS_600
Synchronically felt foreignness of individual constituents	Coded manually	OHS_600
Compound spelling variant in the earliest OED attestation	Coded manually based on the OED	OHS_600
Age of the compound	Coded manually based on the earliest attestation in the OED	OHS_600
Spelling development of the compound	Coded manually based on the OED and the spelling in OHS_600	OHS_600

Table 5.98 (*cont.*)

VARIABLE	METHOD	SAMPLE
I) Discourse variables		
Editing	Spelling-sensitive CompSpell search in corpora of edited vs. unedited language (BEO6, Blog Authorship Corpus)	Master List
Variety	Spelling-sensitive CompSpell corpus search in corpora of British vs. American English (LOB/ FLOB vs. Brown/Frown)	Master List
J) Systemic variables		
Spelling-sensitive left constituent family size	CompSpell search in MED	OHS_600
Spelling-sensitive left constituent family frequency in BNCwritten	Copying of external automated search results for the frequencies of left constituent family members in BNCwritten into the database	OHS_600
Spelling-sensitive right constituent family size	CompSpell search in MED	OHS_600
Spelling-sensitive right constituent family frequency in BNCwritten	Copying of external automated search results for the frequencies of right constituent family members in BNCwritten into the database	OHS_600
K) Extralinguistic variables		
Speed of typing	Spelling-sensitive CompSpell search in blog vs. chat corpora (Blog Authorship Corpus, NPS Chat Corpus)	Master List
Spatial restriction	Spelling-sensitive CompSpell search in blog vs. text message corpora (Blog Authorship Corpus, CorTxt Corpus)	Master List

that were coded for the individual samples and of the method used in each case.

The preceding sections tested to what extent variables from different domains correlate with compound spelling preferences in British English, with a focus on established biconstituent compounds that are spelled identically in all five to six dictionaries in which they occur. This subgroup was selected as a central and prototypical sample. While some of the variables under consideration did not prove relevant to the spelling of English compounds (e.g. silent <e> at the end of the first constituent; cf. Hypothesis D1), the overwhelming majority show a strong correlation and can most probably be considered determinants of variant selection and not merely accidental correlations. As a consequence, this is presumably what renders the spelling of English compounds so difficult for language users: not the lack but the overabundance of underlying principles, many of which escape conscious perception. In view of the large number of variables with a significant effect on the spelling of English compounds, one may consider speaking of *complexity-induced variance* as a situation with clear rules (in compound spelling, relatively clear rules) which are too complex for the majority of language users to observe in their entirety (cf. Gallmann 2004: 41).

Summaries at the end of each of the subgroups A to L give an overview of the findings for each category. In addition, Table A.9 in the Appendix lists all variables that statistically correlate with open, hyphenated and/or solid spelling for the OHS_600 compounds and can therefore be considered essential for spelling variant selection.

Table 5.99 *Extract from Table A.9*

	Variable	Explanation	Value	O	H	S	Pref.
A2	Ident_lett_r	Repeated letters across boundary	+	+	+	--	2
B2a	Syll_total_r	Number of syllables	2	-	0	+	1
B2b	Syll_total_r	Number of syllables	3	+	0	-	1
B2 c	Syll_total_r	Number of syllables	4 or more	++	0	--	1+
B8a	Syll_1_r	Length of first constituent (syllables)	1	-	+	+	2
B8b	Syll_1_r	Length of first constituent (syllables)	2 or more	++	-	--	1+

For a quick overview of the detailed discussions in Chapter 5, the code in the first column of Table A.9 refers to the hypothesis at which the respective variable was tested (e.g. *A2*). Several types of effect linked to a single hypothesis are further distinguished by letters (e.g. *B2a*, *B2b* and *B2c*). Columns 2 to 4 indicate the variable's code, paraphrase its meaning and point out the value to which the spelling tendencies in columns 5 to 7 apply. Thus a plus sign in the value column indicates the presence of the quality described in column 2 (such as repeated letters across the constituent boundary for *A2*), whereas *B2a*, *B2b* and *B2c* refer to compounds comprising two, three and four or more syllables, respectively. The codes in columns O, H and S indicate the strength of the tendencies towards open, hyphenated and solid spelling; i.e. that the respective spelling variant is:

- ++ far more frequent than expected by chance
- + more frequent than expected by chance
- o about as frequent as expected by chance
- less frequent than expected by chance
- far less frequent than expected by chance.

The allocation of these codes depends on the relative proportions for each variable, but some generalisations are possible: *o* applies to proportions of about 33 per cent (± 5 per cent). If the proportions of two spelling variants differ by a maximum of about 5 per cent, they are marked with the same sign (e.g. + in the case of *Syll_1_r*), particularly if the absolute numbers are small. The classification as ++ corresponds roughly to a proportion of two-thirds or higher. The last column in Table A.9 documents a preference for either one or two of the three spelling types. It indicates how much the respective variable (in combination with a particular value) contributes to the distinction between the three spelling variants. A particularly strong preference for one of the three spelling variants is marked with a plus sign, e.g. for open spelling in *B2c*, which is classified as ++ compared to the values *o* and -- for hyphenation and solid spelling.

In addition to investigating the potential determinants of English compound spelling variant selection summarised in Table A.9, the preceding sections tested a number of related hypotheses with the following results:

- Using the spelling in the first OED attestation, it was possible to contradict the general view that compounds start their life open, then go through a hyphenated stage and finally become solid – at least for the subgroup of compounds under consideration.

- For the same sample, spelling differences between present-day British and American English are also smaller than expected from the literature, and hyphenation is not more common in British than in American English.
- Similarly unexpectedly, reduced readability across the constituent joint (e.g. due to consonant clusters or garden path clusters) does not disfavour solid spelling.
- As far as linguistic context is concerned, compounds which are not exclusively spelled solid in the dictionaries have a very slight tendency to be hyphenated before a noun and to be spelled open in other contexts, but this result is not statistically significant.
- With regard to extralinguistic variables, we find that spatial restriction does not favour solid spelling.
- Higher speed of typing, by contrast, favours open spelling indeed.
- Furthermore, and as expected, edited corpus texts follow the spelling in the dictionaries more closely than unedited texts.
- On a more general level, we also find that compounds with heterogeneous or complex constituents disfavour solid spelling.

Assuming that the data and methods used yielded representative results for the current state of British English, we can conclude from all of this that the spelling of English compounds is not as chaotic as has been believed in the past, but that it is actually influenced by an exceedingly large number of variables. Note also that 7,970 and thus 86 per cent of the 9,258 biconstituent Master List compounds are spelled identically in all the dictionaries in which they occur (5,118 exclusively open, 987 exclusively hyphenated and 1,865 exclusively solid), which means that the difference between the various dictionaries is small with respect to compound spelling, possibly due to the general use of corpora in modern lexicography. While there are no binding rules regarding the spelling of English compounds (cf. Chapter 3), the present study's results show that there is a relatively agreed-upon way of spelling a large number of English compounds, so that some spellings can be considered less usual than others. Juhasz, Inhoff and Rayner's (2005: 294) claim that spacing in English compounds "can be manipulated without violating orthographic conventions" therefore needs to be refuted: even if a spelling such as *?agony aunt* does not go beyond the options provided by the system of English orthography, it violates the conventions in the sense of a norm (cf. Coseriu 1978: 44).

The foregoing sections also discuss the findings of the present study in relation to the findings of previous research. Where these differ (e.g. with

regard to high compound frequency; cf. Hypothesis C1), the reasons can frequently be found in a major difference in research design, namely the restriction to nominal noun+noun compounds from corpora in the literature as against the application of a broad compound concept comprising all possible part-of-speech combinations in the present study. To what extent the results obtained here can be applied to other compounds (less established, novel or with more than two constituents) is discussed in the following.

PART III

Modelling English Compound Spelling

The previous chapters laid the theoretical foundations for the empirical study of English compound spelling and presented the results of a comprehensive empirical study. In the following chapters, this wealth of data is first used to establish compound spelling algorithms on various levels of complexity, thereby representing an applied perspective with possible pedagogical focus. The subsequent sections are characterised by a move back to the theoretical level and attempt to provide a model of English compound spelling which takes into account both the present state and its inherent dynamics towards linguistic change.

Compound Spelling Heuristics

While many English compounds are spelled automatically (without the spellers' conscious awareness whether they are using open, hyphenated or solid spelling), the spelling of other compounds represents an instance of uncertainty in decision-making (cf. Kahnemann and Tversky 1982: 512), with language users stating explicitly that they find spelling variant selection difficult. Since "outcome feedback is the main source of information for evaluating the quality of our decision/judgment rules" (Einhorn 1982: 269), it is difficult for language users to sharpen their intuition regarding English compound spelling, as they usually get little feedback on the appropriateness of their choices (with the possible exception of the school context). Since language is not stable but subject to change, linguistic intuition may need to adapt to changing usage from time to time (cf. 7.5) – in contrast to more permanent types of intuition, e.g. that for predicting the trajectory of a ball in order to catch it (cf. Gigerenzer 2008: 17–19).

When language users feel uncertain regarding the decision for a particular linguistic form, they can either attribute this to the external world (e.g. by assuming that the spelling of English compounds is chaotic) or to their own state of knowledge (e.g. by assuming that there are orthographic principles but that they do not know them or have forgotten them). If the problem cannot be solved introspectively by looking for a spelling which seems familiar and intuitively looks right "in a reasoned mode" (Kahnemann and Tversky 1982: 519), possible decision-making strategies are

- avoiding the problem (e.g. by using a paraphrase instead of the compound)
- alternating between variants (which contradicts the generally advocated principle of consistency in spelling; cf. e.g. Ritter 2005b: xi)
- using a combination of two variants (but juxtaposed alternatives like *girl(-)friend* are extremely uncommon outside metalinguistic use)

- using an intermediate variant as a compromise (but the exploitation of the distance between the constituents in solid vs. open spelling is restricted to handwriting)
- seeking advice (e.g. in an up-to-date dictionary; cf. Clark 1990: 189)
- guessing
- using some kind of algorithm (like those discussed in what follows).

If all spelling variants were generally acceptable for every compound (which they are not), 100 per cent of all choices would be correct. If compounds were spelled by chance and only permitted one acceptable spelling variant each, then open, hyphenated and solid spelling should occur in about 33.3 per cent of cases each (but this is not the case either). A higher prediction accuracy than that of random guessing (namely 67 per cent for the Master List compounds, 70 per cent for CompText types and 64 per cent for CompText tokens) could be achieved by selecting open spelling as the most frequent spelling of biconstituent compounds by default. However, the sole use of the most frequent spelling variant has the important disadvantage that 100 per cent correct predictions for true open compounds are counterbalanced by 0 per cent both for true hyphenated and for true solid compounds.¹ Since an ideal algorithm should not completely sacrifice prediction accuracy in one or two categories for higher predictability in the other, it makes sense to look for possible ways of improvement that can do justice to all three categories.

As pointed out in the introduction, one of the aims of the present study was to determine tentative spelling strategies for biconstituent British English compounds in the early twenty-first century for a possible pedagogical application (e.g. for learners of English or less confident spellers). This reverses the perspective of the empirical study (which regards the spelling as the dependent variable and determines how characteristic the features are for each spelling variant) by investigating to what extent a compound's inherent features can serve to predict an acceptable present-day spelling for that compound.

6.1 Method

In the following, several heuristics are considered with the aim of raising prediction accuracy. Different measures and methods can be used for the

¹ In the context of the present study, *true* corresponds to the spelling variant favoured by at least the majority (if not all) of the dictionaries. More generally, it refers to a variant which is accepted as correct by the majority of language users.

determination of such heuristics, e.g. odds ratios, discriminant analysis (cf. e.g. Gries 2003a: 164) or variable rule analysis (cf. e.g. Tagliamonte and Baayen 2012). Since all of these have their advantages and disadvantages, it was decided to use decision trees, because their results can be communicated most easily to human users of spelling strategies while simultaneously permitting a computer-generated output. The decision trees were produced with the *party* package in R (cf. Hothorn, Hornik and Zeileis 2006). Only those context-independent variables (which excludes speed, spatial restriction, grammatical context and editing) that were significant for the OHS_600 dataset in the testing of the hypotheses earlier (cf. Table A.9 in the Appendix)² were used in the determination of the most comprehensive compound spelling algorithm:

In a stepwise procedure, various types of feature were omitted with the aim of arriving at a strong, simple and maximally suitable algorithm for pedagogical application:

- **Subjective variables:**
 - idiomaticity
 - synchronically felt foreignness
 - main stress
 - part of speech of the first constituent (which is determined using an imagined prototypical context)
 - part of speech of the second constituent (cf. earlier)
- **Variables requiring the use of a reference work:**
 - age of the compound
 - earliest attested spelling in the OED
 - combination of Germanic and Romance constituents
- **Variables requiring a certain amount of linguistic training:**
 - presence of one or more complex constituents
 - presence of non-compound-final lexical suffix

² This approach has two potential weaknesses (pointed out by Antony Unwin): on the one hand, it may include factors which are significant on their own but actually highly correlated, so that the inclusion of the second (or third) factor has no clear effect beyond that of the first factor. However, that is no problem, since the program R, which was used for the generation of the decision trees, simply ignores the correlated variables. On the other hand, it may also be that variables which are not significant on their own are actually important (because they become important in combination with other variables, are very important for a very small subset of compounds or were underrepresented in the data for some other reason). While this certainly imposes restrictions on the possible results that can be achieved automatically by the present study's algorithms, a human speller might also consider the list of spelling tendencies without statistical backing (cf. Table A.10 in the Appendix) in addition to the algorithms.

Table 6.1 *The comprehensive Algorithm 1 comprising all significant variables (initial version)*

o_1_CC_2_r	Consonant cluster across boundary
Ident_lett_r	Repeated letters across boundary
o_1_VV_2_r	Vowels across boundary
Syll_total_r	Number of syllables
Lett_total_r	Number of letters
Syll_diff_12	Length difference between constituents exceeds 1:2 (syllables)
Lett_diff_12	Length difference between constituents exceeds 1:2 (letters)
Lett_1_r	Length of first constituent (letters)
Lett_2_r	Length of second constituent (letters)
Syll_1_r	Length of first constituent (syllables)
Syll_2_r	Length of second constituent (syllables)
Total_BNC_r	Total BNCwritten frequency of all spelling variants
Freq_rvs2	Combined frequency ranges of first and second constituent
Freq_diff_12	Frequency difference between first and second constituent
Freq_1_r	Frequency range of first constituent
Freq_2_r	Frequency range of second constituent
Stress	Main stress
Complex_const_r	One or more complex constituents
Nonfin_lex_suff	Non-compound-final lexical suffix
Final_ingeder_r	Compound-final suffix <i>-ing</i> , <i>-ed</i> or <i>-er</i>
Morphol_struct_r	Head-final morphological structure
PoS_comp_r	Part of speech of compound
PoS_1	Part of speech of first constituent
PoS_2_r	Part of speech of second constituent (adv + v together)
Lexgr_12_diff	Combination of lexical and grammatical constituents
Idiom	Idiomatcity
Age_r	Age of compound
Earl_spell_r	Earliest attested spelling in OED
Mixed_etym	Germanic + Romance constituents
Foreignness	Synchronically felt foreignness
LS_r	Spelling preferred by left constituent family size
RS_r	Spelling preferred by right constituent family size
LF_r	Spelling preferred by left constituent family frequency
RF_r	Spelling preferred by right constituent family frequency

• **Variables requiring explicit frequency data:**

- combined frequency ranges of first and second constituents
- frequency difference between first and second constituents
- frequency range of the first constituent
- frequency range of the second constituent
- spelling preferred by left constituent family size
- spelling preferred by right constituent family size

- spelling preferred by left constituent family frequency
- spelling preferred by right constituent family frequency.

This stepwise reduction permitted testing various types of algorithm, such as a user-friendly algorithm (including the significant features which are easy to apply for language users without prior linguistic training), a take-the-best heuristic or a maximally efficient algorithm. The tests were carried out with the *party* package in R (cf. Hothorn et al. 2006) using the function *tree* for the creation of conditional inference trees. The default parameters used were *teststat* = quad; *testtype* = Bonferroni; *mincriterion* = 0.95; *minsplit* = 20; *minbucket* = 7. Nodes needed to contain a minimum of seven items and splitting was only considered if a node contained at least twenty items. A Bonferroni approach was used to control for multiple tests of significance. The performance of the algorithms was judged from the proportion of correct predictions for different sets of data:

- A) The training set OHS_600. The algorithms were considered to predict the spelling correctly if their prediction coincided with the form used unanimously by the reference works.
- B) Since the algorithms are based on biconstituent compounds which occur in many dictionaries and are always spelled identically (OHS_600), they were tested on a highly similar dataset, namely the OHS_extra compounds. These comprise all compounds from the Master List corresponding to the same criteria as the OHS_600 compounds, but which were not selected for OHS_600 in the randomised process. The algorithms were considered to predict the spelling correctly if their prediction coincided with the form used unanimously by the reference works.
- C) In order to check whether the algorithms also predict the spelling of biconstituent compounds which are recorded in many dictionaries but offer a certain degree of variation, a subset of 764 biconstituent compounds was extracted from Master_5+: Master_5+_tendency comprises all compounds occurring at least five times in the dictionaries with
 - a. a majority of open and some hyphenated spellings (Oh)
 - b. a majority of open and some solid spellings (Os)
 - c. a majority of hyphenated and some open spellings (Ho)
 - d. a majority of hyphenated and some solid spellings (Hs)
 - e. a majority of solid and some open spellings (So)
 - f. a majority of solid and some hyphenated spellings (Sh).

The unmentioned spelling variant always has a value of zero. The algorithm was considered to predict the spelling correctly if its prediction corresponded to the spelling favoured by the majority of the dictionaries.

- D) Since it was assumed that the heuristics derived for established compounds with unanimous spelling in reference works can also be applied to more recent, unestablished compounds, yet another test was carried out on the biconstituent compounds from the CompText corpus (cf. 4.2) which are not included in the Master List. The algorithm was considered to predict the spelling correctly if its prediction coincided with the form used in the corpus. This was possible, since the corpus contained no contradictory spellings for the same compound with the same part of speech.

The test compounds from the corpus were selected manually based on the criteria described in Section 2.6 and usually labelled according to the present study's part-of-speech conventions.³ Since compounds may be contained within compounds (with possible spelling differences by comparison to isolated occurrence), the longest sequence accepted as a compound was always retained (e.g. *banner+head+lines* rather than *head+lines* only). Consecutive compounds which do not combine into a larger compound were recorded separately, e.g. *ship+crowded* and *water+front* in the sequence *ship-crowded waterfront*. The resulting compounds were lemmatised and deleted if they also occurred in the Master List with the same part of speech.⁴ Since one may assume that the spelling of proper names is usually determined by some group – such as the founders of a brand, or the community naming a street – it makes little sense to devise

³ Since the corpus texts contain several compound numerals (e.g. *four thousand*), *num* was introduced as an additional part-of-speech label. *Another* was classified as a compound determiner (*det*). The large number of proper nouns as compounds and constituents and the presence of some proper adjectives (e.g. *English*) made it seem appropriate to distinguish their effect from that of general nouns and adjectives by introducing the part-of-speech codes *n_pr* and *adj_pr*.

⁴ Seventy-two CompText compound types are contained in at least one dictionary with the corresponding part of speech. Most of the compounds from the corpus texts are spelled identically in all the (usually five or six) dictionaries in which they occur: twenty-two are always open, seven are always hyphenated and thirty-one are always spelled solid. From this, we can conclude that the spelling of English compounds in edited texts is indeed very close to that advocated by reference works (cf. also Hypothesis I2) – be it because professional authors and publishers check dictionaries in case of doubt or because lexicographers use edited texts to decide upon the spellings recorded in dictionaries.

spelling strategies for names. The large number of proper lexemes was therefore disregarded in the testing of the spelling heuristics. This leaves us with 198 new compound types in CompSpell. Some of these (e.g. *points swap scam* or *language guru*) refer to rare phenomena, whereas a few others seem general in their scope (e.g. *sale price* or *membership terms*).⁵ One hundred fifty-four of the compounds from that group have two constituents; forty have three (e.g. *fair+ground+booth* or *red+carpet+ready*), three have four (e.g. *make+up+staying+power*) and one compound has five constituents: *NOT WANTED ON VOYAGE stickers*, which uses capitalisation as a spelling device. For technical reasons, three biconstituent compounds containing figures (*19th+century*, *1.8+million* and *26+year*) and two compounds including hyphenated prefixations (*pre-awards drink* and *pre-match debate*) had to be excluded from the testing procedure, which yielded 149 compounds for the testing.

6.2 Results

In the discussion of the test results for the algorithms that follows, one should not overlook the fact that the present study differs from many others in that it does not consider a binary decision but as many as three alternatives. The prediction accuracy of the algorithms must therefore be expected to be numerically inferior to that achieved in studies distinguishing merely between two alternatives (cf. e.g. Gries 2003a: 165–166, with a prediction rate of 83.1 per cent for particle placement, or the various results for compound stress reported in Plag 2010: 253–254).

The comprehensive Algorithm 1 was found to predict the spelling correctly (cf. the shaded cells in Table 6.2) for 511 and thus 85.2 per cent of all OHS_600 compounds based on the majority spelling in the corresponding final node of the decision tree (cf. Figure 6.1). Since OHS_600 is the only sample for which all variables had been coded, the comprehensive algorithm could only be tested on that sample, so that the results in Table 6.2 can merely indicate the extent to which the algorithm explains the training set.

Figure 6.1 represents the detailed decision tree. Two variables occur in two different places, namely stress [Stress] and part of speech of the first

⁵ Some of these ‘new’ CompText compounds are actually contained in LDOCE but were not coded as compounds by the lexicographers and are therefore not included in the Master List (e.g. *anything* or *crafts+man*).

Table 6.2 *Predictive accuracy of the comprehensive Algorithm 1 (all significant variables) for OHS_600*

		Actual spelling		
		O	H	S
Predicted spelling	O	157	7	14
	H	3	187	19
	S	40	6	167

Table 6.3 *Variables retained in the final version of the comprehensive Algorithm 1*

Syll_total_r	Number of syllables
Lett_2_r	Length of second constituent (letters)
Stress	Main stress
Final_ingeder_r	Compound-final suffix <i>-ing</i> , <i>-ed</i> or <i>-er</i>
Morphol_struct_r	Head-final morphological structure
PoS_comp_r	Part of speech of compound
PoS_1	Part of speech of first constituent
PoS_2_r	Part of speech of second constituent (adv + v together)
Earl_spell_r	Earliest attested spelling in OED
LS_r	Spelling preferred by left constituent family size
RS_r	Spelling preferred by right constituent family size

constituent [PoS_1]. The algorithm only uses eleven of the thirty-four original significant variables (cf. Table 6.1), because all others were ignored in the automatic computation of the algorithm due to their correlation with other variables. Conveniently, three variables which are problematic in some respect, namely o_1_CC_2_r (which makes counter-intuitive predictions; cf. 5.1.1.1), and Syll_diff_12 and Lett_diff_12 (whose results slightly contradict each other; cf. 5.2.9) were automatically excluded.

It is striking that the automatically reduced set completely lacks any variables related to spelling and semantics. Furthermore, in spite of the fact that frequency is the most important predictor for the spelling of noun+noun compounds in Kuperman and Bertram (2013: 959), frequency information merely enters the decision tree via right and left constituent family size in nodes 14 and 18 on the lowest level of the decision tree and with weak significance values compared to the other nodes. In view of the important role that frequency usually plays in language (cf. 7.3), a possible explanation might be that frequency correlates very strongly with one or



Table 6.4 *Correlation between compound length (in syllables) and frequency (in BNCwritten) for OHS_600*

			Syll_total_r			Total
			2	3	4–9	
Total_BNC_r	low	Count	74	51	28	153
	(0–17)	% within Syll_total_r	23.4%	26.4%	30.8%	25.5%
	mid	Count	145	103	47	295
	(18–126)	% within Syll_total_r	45.9%	53.4%	51.6%	49.2%
	high	Count	97	39	16	152
	(127–55,000)	% within Syll_total_r	30.7%	20.2%	17.6%	25.3%
Total		Count	316	193	91	600
		% within Syll_total_r	100.0%	100.0%	100.0%	100.0%

more of the retained variables – most probably length. This hypothesis was tested by carrying out Pearson’s chi-square test for the dependent variable ‘frequency’ (more specifically [Total_BNC_r], i.e. recoded frequency comprising all spelling variants in BNCwritten) and the independent variable ‘compound length’ (more specifically [Syll total], i.e. the number of syllables of the compound). As expected, one can observe a statistically significant correlation between compound length and corpus frequency ($p = 0.028$) for the OHS_600 compounds – even though it is weaker than expected. Regardless of compound length, intermediate frequency is most common – but if we consider the second most frequent value in each length category, short compounds tend to be more frequent (30.7 per cent), and long compounds tend to be less frequent (30.8 per cent). This result also justifies the omission of frequency data from the simplified algorithms discussed later.

The user-friendly Algorithm 2 consists of merely seven variables (cf. Table 6.5) and performs almost as well as the comprehensive algorithm for the OHS_600 compounds (cf. the shaded cells in Table 6.6): with 488 correct predictions and a prediction accuracy of 81.3 per cent, it is merely 3.9 per cent below the prediction rate of Algorithm 1 for that sample.

Testing with other reduced sets of variables confirmed the outstanding role of part of speech (more specifically, grouped part of speech of the compound [Pos_comp_r]) for the spelling algorithms. This result is in line with Sepp (2006: 97), who observes that “the number of syllables is the strongest single predictor of the closed and open forms” of noun+noun

Table 6.5 *The user-friendly Algorithm 2*

Syll_total_r	Number of syllables
Lett_1_r	Length of first constituent (letters)
Lett_2_r	Length of second constituent (letters)
Final_ingeder_r	Compound-final suffix <i>-ing</i> , <i>-ed</i> or <i>-er</i>
Morphol_struct_r	Head-final morphological structure
PoS_comp_r	Part of speech of compound
Lexgr_12_diff	Combination of lexical and grammatical constituents

Table 6.6 *Predictive accuracy of the user-friendly Algorithm 2 for OHS_600*

		Actual spelling		
		O	H	S
Predicted spelling	O	136	10	11
	H	4	185	22
	S	60	5	167

compounds. Since part of speech occurs at the top of the decision tree in the comprehensive Algorithm 1 (cf. Figure 6.1) and performs best in discriminating between the three spelling variants in the training set OHS_600, its predictive accuracy as a maximally simple take-the-best heuristic was tested. Different studies (e.g. Gigerenzer and Goldstein 1999; Czerlinski, Gigerenzer and Goldstein 1999) have shown that take-the-best performs better in predictions than more complex strategies, such as multiple regression. While “the complex strategy can weigh its many reasons so that the resulting equation fits well with what we already know”, only “part of the information is valuable for the future”, and “the art of intuition is to focus on that part and ignore the rest”, so that a “simple rule that relies only on the best clue has a good chance of hitting on that useful piece of information” (Gigerenzer 2007: 85). Thus a simple temperature curve is likelier to predict next year’s temperatures correctly than one which matches this year’s data perfectly (Gigerenzer 2007: 151), because the latter takes into account irrelevant effects which cannot be generalised for the future (Gigerenzer 2008: 162). Czerlinski, Gigerenzer and Goldstein

Table 6.7 *Predictive accuracy of the take-the-best Algorithm 3 for OHS_600*

		Actual spelling		
		O	H	S
Predicted spelling	O	200	45	181
	H	0	155	19
	S	0	0	0

(1999: 109) explain such overfitting by analogy with a cyclist who always trains on the same track and who “may get so used to this pattern that he can no longer deal well with other combinations of hills and plains”. A large number of cues will cause accidental correlations; if these are built into the model, they lead to inaccurate predictions (Czerlinski et al. 1999: 109), whereas simpler strategies are more robust in generalization (e.g. for the prediction of high school dropout rates; cf. Czerlinski et al. 1999: 109). A further advantage of one-reason decision-making is that it is much faster than more sophisticated complex strategies (Gigerenzer 2008: 256). This is of particular interest to spelling, because on the subconscious level, models which require large amounts of processing power – such as weighted linear models or multiple regression – are unlikely to have any psychological validity (Gigerenzer and Goldstein 1999: 84).

However, the take-the-best strategy does not seem to work so well for English compound spelling: Algorithm 3, which bases variant selection exclusively on the compounds’ part of speech, achieves a predictive accuracy of merely 59.2 per cent (cf. Table 6.7) and never predicts solid spelling (neither correctly nor incorrectly). This shows the necessity of combining this powerful variable with others in order to achieve better and more balanced results.

In view of the intended pedagogical application, a convenient spelling heuristic should aim at a compromise between large prediction accuracy and a small number of features. Using the user-friendly Algorithm 2 as the point of departure for a maximally efficient spelling algorithm, the number of features was reduced, e.g. by ignoring anterior splits which did not change the result for the relevant final node (like node 5 in Figure 6.1, because solid spelling dominates in nodes 6 and 7 anyway) or final nodes affecting only a very limited number of compounds (like node 12 in Figure 6.1, which only concerns 9 out of 600 compounds).

The reduction of the number of features finally resulted in Algorithm 4, which comprises only three variables: the part of speech of the compound [PoS_comp_r], the number of syllables of the compound [Syll_total_r] and the number of letters of the second constituent [Lett_2_r]. Figure 6.2 shows the decision tree achieved if mincriterion is set at 0.99999 in order to reduce the number of nodes.

With a maximum of only three steps, this algorithm can be applied very easily by human users. Its prediction accuracy of 80.7 per cent for the OHS_600 compounds (cf. Table 6.8) would seem to make it the simplest algorithm with reasonable prediction accuracy that can be obtained from the data. Even though [Lett_2_r] only affects twenty-three OHS_600 compounds, it was retained due to its efficiency, with 100 per cent prediction accuracy for hyphenation.

Since the aim of a heuristic is the application to new sets of data, the prediction accuracy of Algorithm 4 (also referred to as the CompSpell algorithm) was tested for various types of sample (cf. 6.1). However, the fact that there are no grammatical compounds in OHS_600 (as the training set for the algorithms) means that these could not be considered in the testing and had to be removed prior to testing from OHS_extra (6), Master_5+_tendency (2) and CompText (16).

Algorithm 4 predicts 2,944 out of 3,858 (= 76.3 per cent) OHS_extra compounds correctly. Since this value lies 4.4 per cent below the prediction accuracy for the OHS_600 compounds (which were randomly selected from the same larger sample as the OHS_extra subset), one may conclude that the algorithm is slightly over fitted to the training set but still relatively efficient.

Algorithm 4 predicts 465 out of 762 (= 61.0 per cent) Master_5+_tendency compounds correctly. The low proportion of correct predictions can be explained by the fact that the compounds in this sample are less clear in their spelling preferences, since the dictionaries disagree among

Table 6.8 *Predictive accuracy of the maximally efficient Algorithm 4 for OHS_600*

		Actual spelling		
		O	H	S
Predicted spelling	O	155	14	30
	H	0	178	19
	S	45	8	151

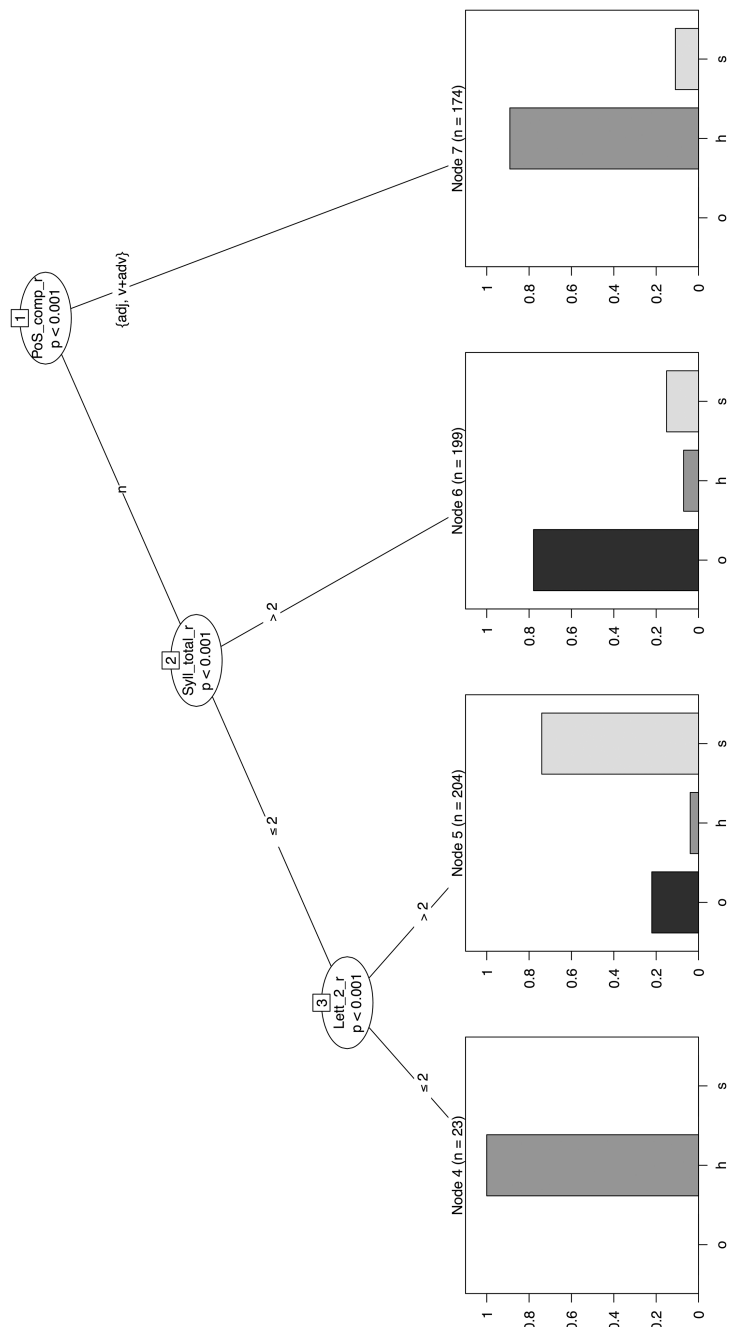


Figure 6.2 Decision tree for the maximally efficient Algorithm 4 for OHS_600

each other. As a consequence, the result of 61.0 per cent refers to the correctly predicted majority spellings only and ignores the fact that a second spelling variant is considered correct by part of the dictionaries and thus probably also by many language users. Therefore, a certain number of spellings outside the shaded cells marking correct predictions in Table 6.10 would actually be considered appropriate in actual language usage.

Table 6.9 *Predictive accuracy of the maximally efficient Algorithm 4 for OHS_extra without grammatical words*

		Actual spelling		
		O	H	S
Predicted spelling	O	1723	33	166
	H	17	278	110
	S	574	14	943

Table 6.10 *Predictive accuracy of the maximally efficient Algorithm 4 for Master_5+_tendency without grammatical words*

		Actual spelling		
		O	H	S
Predicted spelling	O	109	37	40
	H	15	177	26
	S	143	36	179

Table 6.11 *Predictive accuracy of the maximally efficient Algorithm 4 for CompText without grammatical words*

		Actual spelling		
		O	H	S
Predicted spelling	O	76	3	3
	H	3	15	11
	S	13	3	6

Algorithm 4 predicts 97 out of 133 (= 72.9 per cent) CompText compounds (which are not listed in LDOCE) correctly. Note, however, that this test is less informative than the others, since it usually relies on a single occurrence of each compound (which raises the probability of idiosyncratic spellings by comparison to the compounds from dictionaries).

Not surprisingly, the highest prediction accuracy among all test sets (76.3 per cent) was achieved for the OHS_extra compounds (which are most similar to the training set), and the lowest prediction accuracy (61.0 per cent) was achieved for Master_5+_tendency (the sample with the strongest degree of variation). The CompText compounds can be considered a random sample with an unclear amount of spelling variation. This is reflected in the algorithm's predictive accuracy for this sample, which lies between that for the other two samples, with 72.9 per cent correct predictions. Across all test samples, the most common type of error is the incorrect prediction of solid spelling for compounds that are better spelled open. As a consequence, human users of the algorithm are well advised to select open rather than solid spelling if they have doubts about the output of the algorithm.

Note that the absence of grammatical words from the training set OHS_600 means that the algorithm cannot spell such words and treats them as errors. While this may seem like a very important drawback at first sight, the practical consequences are less dramatic: there are few compounded grammatical words in the data except for a few high-frequency items such as *in+to* in the CompText corpus, but these are unlikely to pose any problems to language users precisely because they are so frequent and established. Since grammatical words are under-represented in the Master List and since the analyses in Section 5.6.1.6 are inconclusive, a complementary analysis was carried out on the pronouns, articles, conjunctions, prepositions and interjections retrieved by an advanced search in the OED as the largest reference work for British English. Only those items were retained which correspond to the compound definition followed in the present study and which are in use in standard present-day British English (cf. Table A.11 in the Appendix). Based on this single-source sample of limited representativeness, it is possible to derive the following tentative strategy for grammatical English compounds: solid spelling is always correct for conjunctions and pronouns, even if they consist of more than two constituents (e.g. *what+so+ever*), and also for the majority of prepositions. About twice as many interjections are hyphenated rather than spelled open (particularly when the two parts are similar or identical),

Table 6.12 *The maximally efficient Algorithm 4a for all parts of speech*

• Adjective (<i>broken-down</i>)	Hyphenated
Adverb (<i>well-nigh</i>)	
Verb (<i>chain-smoke</i>)	
Numeral (<i>twenty-two</i>)	
Interjection (<i>night-night</i>)	
• Grammatical word (<i>into</i>)	Solid
• Noun	
• three or more syllables (<i>bathing suit</i>)	Open
• two syllables	
○ second constituent: up to two letters (<i>close-up</i>)	Hyphenated
○ second constituent: more than two letters (<i>coastline</i>)	Solid

but solid spelling is very rare in this group. While the OED does not list compound numerals as lemmas in their own right (except if they stand for a different concept, such as *twenty-two* for a twenty-two calibre rifle), these seem to be hyphenated in other sources. If the maximally efficient Algorithm 4 is combined with these findings for the grammatical words, numerals and interjections, we obtain the maximally efficient Algorithm 4a for all parts of speech (cf. Table 6.12).

Algorithm 4a could not be tested in the context of the present study, but as only a small number of compounds causing spelling difficulties are grammatical words, spelled-out numerals or interjections, the results should be very similar to those for Algorithm 4 and predict a commonly acceptable spelling for roughly three out of four compounds. While previous accounts of English compound spelling mention individual parts of the pedagogical algorithms given earlier in different texts and places (cf. Chapter 5), their combination in this reduced number and sequence in the present study represents an innovation.

6.2.1 Native Speaker Study

While the foregoing results are interesting on their own and suggest a relatively good fit of the model to the data, the question remains how to evaluate the performance of a model with prediction accuracies ranging from 61.0 per cent for the majority spelling of compounds with several acceptable spelling alternatives to 76.3 per cent for relatively clear cases and 80.7 per cent for the training set. Compared to the potential upper limit of 100 per cent, these proportions seem modest – but on the other hand, they

Table 6.13 *Example test item for the native speaker study*

1	2	3	4	5
girl		friend	n	

are an impressive improvement on random guessing with a predictive accuracy of about 33 per cent (if only one spelling per compound was acceptable). As a consequence, it makes sense to compare the performance of the model to that of human spellers with high linguistic competence. To this end, two native speakers of British English in their thirties to forties (one female, one male) were asked to indicate the most likely spelling for all OHS_600 compounds. Since both subjects had an academic background and had been teaching English as a foreign language for a considerable number of years, it was expected that they could not merely rely on normal native speaker intuition but were also highly aware of normative issues regarding spelling.

The subjects were given an electronic version (*Word* or *Excel*, according to their preferences) of the test items in alphabetical order, together with the written instruction to decide upon the spelling of the compounds very quickly, following their intuition and without checking a dictionary or asking anyone. They were also told that there was no right or wrong answer, and they were not informed about the existence of any spelling tendencies in the data. The compounds were presented in the following format:

Column 1 contained the first constituent, column 3 the second constituent and column 4 the part of speech of the compound. While the approach could possibly be improved by creating a full-sentence context for each of the compounds, the fact that both participants were teachers of English permitted the use of part of speech as a shortcut. The participants were asked to insert what they believed to be most likely into column 2: a hyphen for hyphenation, nothing for solid spelling and a space or some symbol of their choice for open spelling. In addition, they were encouraged to mark compounds which they found difficult to spell in column 5 by inserting a hash key or some other symbol. While more detailed testing using additional methods would be desirable, this small-scale test functions as a legitimate control for the performance of the algorithm.

Table 6.14 *Proportion of correct predictions by the CompSpell algorithm and two English native speakers for OHS_600 and 100-word samples from OHS_extra and Master_5+_tendency*

	OHS_600	OHS_extra sample	Master_5+_tendency sample
CompSpell algorithm	80.7%	72.0% (total list: 76.3%)	59.0% (total list: 61.0%)
Native speaker 1	74.8%	78.0%	58.0%
Native speaker 2	69.5%	82.0%	55.0%

Participant 1 coincided with the unanimous dictionary spellings for 449 of the 600 OHS_600 compounds and achieved a prediction accuracy of 74.8 per cent. Participant 2 did not indicate any spelling variant for ten compounds and marked them with a question mark as an indication of their difficulty. In the analysis, these items were considered as differing from the spelling favoured by the dictionaries. Altogether, Participant 2 agreed with the unanimous dictionary spellings in 417 of the cases, thereby achieving a prediction accuracy of 69.5 per cent. With 80.7 per cent, the CompSpell algorithm thus performs better than the two native speakers on its training set OHS_600. In view of the large number of compounds in the algorithm's usual test sets, 100 biconstituent compounds each were selected from OHS_extra and Master_5+_tendency by inserting random numbers and retaining the first 100 items after sorting the files. Otherwise, the procedure described earlier was followed.

Participant 1 agreed with the unanimous dictionary spellings for 78 per cent of the 100 OHS_extra compounds; Participant 2 achieved an overlap of 82 per cent. These results are slightly above the 76.3 per cent achieved by CompSpell (which are reduced to 72.0 per cent if only the native speaker study's random sample is considered). Both subjects experienced difficulties with a set of twelve compounds, of which four would have been predicted correctly by the algorithm (two hyphenated adjectives and two open nouns with three syllables).

Participant 1 agreed with the majority spelling of the dictionaries in 58 per cent of the 100 Master_5+_tendency compounds; Participant 2 achieved an overlap of 55 per cent. These results are slightly below the 61 per cent reached by CompSpell. Both subjects experienced difficulties with thirty-one shared compounds, of which ten would have been predicted correctly by the algorithm (three hyphenated adjectives, two

hyphenated verbs, four open nouns with three or more syllables and one noun with a two-letter second constituent).

The native speaker test reveals that even users of English whose linguistic competence can be assumed to correspond to the target level that less confident spellers and advanced learners of English may strive for do not select the unanimous dictionary spelling in about a quarter of the instances. As a consequence, even if the algorithm cannot model variant selection in the reference works completely, we may conclude that the performance of the algorithm is practically identical to that of a native speaker of English with strong linguistic awareness and a high level of competence in language teaching.

Interestingly, the test items marked as difficult by Participant 1 were mostly predicted either correctly or incorrectly both by the native speaker and by the algorithm (e.g. the usually hyphenated and solid nouns *sit-down* and *hard rock* as solid). (Participant 2 did not mark any compounds in the OHS_extra and Master_5+_tendency test sets as difficult.) This result is similar to that obtained by Krott, Baayen and Schreuder (2001: 75) for linking morphemes in Dutch: their model, which also selects the majority variant, achieves results which are similar to those obtained by calculating the average for human subjects – from which they conclude that “[a]pparently, participants and the model find the task equally difficult”. The CompSpell algorithm would have been of little use to Participant 1 if only applied in doubtful cases, but very advanced spellers may profit from the present study’s list of statistically determined exception principles (cf. Table 6.16) to support their intuition.

A very interesting result of the native speaker study is what kinds of compound were spelled differently from the dictionaries. As one can easily observe in Table 6.15, hyphenations were by far the most problematic spelling variant for both participants. Participant 1 chose a different spelling for more than 50 per cent of the 200 compounds which are hyphenated in all the dictionaries. While this could be explained by the fact that this particular native speaker admitted a tendency to avoid hyphens, the proportion is even higher (with more than 60 per cent) for Participant 2, who made no such comment. In addition, fifteen of the nineteen compounds which Participant 1 marked as uncertain with regard to their spelling are exclusively hyphenated in the dictionaries. All this seems to suggest that ‘true’ hyphenations are problematic even for highly qualified language users and that a spelling algorithm which takes them into account is particularly desirable. Incidentally, the CompSpell algorithm performs very well on hyphenations, reaching prediction accuracies of 89.0 per cent

Table 6.15 *Native speaker spellings differing from the unanimous dictionary spellings for OHS_600*

	Dictionary	Spelled as		
Participant 1	O	H	1	22
		S	21	
	H	O	42	110
		S	68	
	S	O	19	19
Participant 2		H	0	
	O	H	13	28
		S	11	
		?	4	
	H	O	101	121
		S	14	
		?	6	
	S	O	28	34
		H	6	

in OHS_600, 85.5 per cent in OHS_extra, 70.8 per cent in Master_5+_tendency and 71.4 per cent in CompText, respectively. All of this considered jointly seems to suggest that the CompSpell algorithm is a satisfactory heuristic to support decision-making, since it is simple and efficient and predicts the most appropriate spelling variant in about three-quarters of instances.

6.3 Discussion

The brief description of the creation and testing of various spelling algorithms is now followed by a more detailed discussion of their theoretical and pedagogical implications.

The spelling of English compounds is highly complex, because many features correlating with variant selection are also correlated among each other, so that it is difficult or impossible to determine their individual effect on compound spelling. For instance, capitalisation is closely linked to

- part of speech (since proper names are usually nouns)
- semantics (since proper nouns are characterised by having reference but no meaning; cf. Lyons 1977: 219)
- word formation type (when the capitalised constituents are acronyms).

The word formation type of the constituents also overlaps with word length, since compounds containing compounds as constituents tend to be longer than others. Part of speech in particular seems to be closely related to several other variables, such as frequency (cf. 5.3.1) or stress (cf. 5.4.4). The maximally efficient algorithm makes use of precisely such correlations in order to arrive at a simpler structure while retaining its predictive accuracy.

Traditionally, compound spelling is often treated in the form of spelling rules like the advice that “[c]ompounds composed of **two nouns** that are **short** and **commonly used**, of which the **first is accented**, are usually written solid” – which combines as many as five prerequisites considering that it occurs in the section on **noun compounds** (Merriam-Webster 2001: 100; highlighting added). The present study, by contrast, formulates its central results in the form of a decision tree structure with several advantages over traditional spelling rules:

- The users of the rule need to combine many variables in their minds to determine whether it applies to the compound they wish to spell. This is cognitively demanding and difficult particularly for younger learners with less explicit formal grammatical training. The spelling algorithm, by contrast, permits an easy linear decision process with one variable at a time. In addition, the ordering of the nodes in the algorithm avoids a conflict between different tendencies for particular combinations of variables and values.
- The spelling process usually requires a different directionality: language users writing a text do not start with the wish of spelling a noun+noun compound with stress on the first of two short and frequent constituents, but with the wish of spelling *bed+room* or *battle+cry*. Style guides and grammars are presumably used infrequently for the spelling of English compounds, since the information can be found more easily and quickly in a dictionary (cf. 1.1.3). However, dictionaries have the disadvantage of listing only established compounds and can thus offer no help regarding very recent or uncommon compounds, whereas the algorithms provide easily retrievable information and can also be applied to new words.
- Furthermore, the indications in this detailed rule are unspecific – for instance, it is unclear what is understood by ‘short’: one syllable, possibly two? Or does shortness refer to the number of letters, and if so, roughly how many? The algorithms presented in the present study, by contrast, are based on relatively clear decisions. Apart from

determining part of speech (which should be clear from the context, particularly in view of the reduction of part-of-speech categories), they only have to count up to three syllables and up to three letters to arrive at a relatively accurate prediction.

- Another disadvantage of the more traditional compound spelling rules lies in their range: if they are very general (e.g. 'spell established compounds solid'), they generate too many exceptions (e.g. adjectives, which tend to be hyphenated). If spelling rules are very detailed, by contrast, they become more difficult to handle, and more rules are necessary to cover all possible cases. For instance, the rule used as an example earlier only applies to nominal noun+noun compounds but provides no guidance regarding the spelling of noun compounds consisting of adjective+noun, or for the spelling of adjective compounds consisting of noun+noun. It is unlikely that all English compounds are covered by the spelling rules assembled in any one style guide or grammar. The algorithm, by contrast, is comprehensive and makes predictions for all possible biconstituent compounds in the English language. While it predicts the spelling of a certain number of compounds incorrectly, the same can be said of many traditional rules in grammars and style guides, because the exceptions they provide can usually only represent a selection.
- Even if there were a grammar describing the spelling of English compounds as a set of both comprehensive and manageable rules, their number would presumably be immense. The algorithm, by contrast, applies the same principle (i.e. a relatively simple decision tree) to all compounds. Particularly when applied repeatedly, it can be memorised with relative ease, and once mastered, the algorithm is always available (even in situations where dictionaries are not) and can be applied more quickly than if spellings had to be looked up.

While one might criticise that the algorithms discussed earlier only make relativistic predictions and link this to general tendencies in the humanities (cf. also Langacker 2008: 88), one should not overlook that the natural sciences also differ in the confidence with which they can make predictions: some laws of physics always act upon matter in the same manner, so that a mechanical process can presumably be predicted with a degree of confidence approaching 100 per cent (provided that the situational characteristics such as mass, surface size, coefficient of friction etc. are known), but many other phenomena escape such precise predictability. For instance, the precise location of an electron within an atom can merely

be located within a probabilistically determined sphere, the atomic orbital (cf. Heisenberg 1927; Vollhardt and Schore 2011: 25), and in the biological sciences, which deal with organisms, one cannot determine accurately in advance which genes are going to be passed on to the next generation.

The weighting and ordering of the variables considered in the present study can be gathered from their location in the decision trees: the closer a variable is situated to the top, the more important its role in spelling variant selection. While some variables only affect very few compounds (e.g. the presence of vowels across the constituent boundaries) and therefore have the value 'zero' for the large majority of compounds, others concern all compounds (e.g. every constituent necessarily has a certain length, frequency and part of speech). While the variables which are only relevant to few compounds are likely to be highly selective (thus all compounds containing acronyms are spelled open), the extremely small number of instances for most of these variables prevents statistical significance. As a consequence, all the variables retained in the pedagogical algorithm are of the comprehensive type.

During the creation of the algorithms, the contribution of individual features to variant selection was also considered. It is now possible to answer the question whether the first or second constituent of a compound plays the more important role in its spelling: since the length of the first constituent is omitted automatically by R and only the length of the second constituent is retained even in the comprehensive Algorithm 1, one may argue that the second constituent of a compound is more influential in this context. This observation is in line with Juhasz et al. (2003), whose study on frequency effects in solid compounds also finds a more important role of the second constituent.

One of the goals of the present study was to explain why language users sometimes find it difficult to select one of the three major variants when spelling an English compound (cf. Chapter 1). A very likely reason is a conflict in variables/values with different directionality: some variables/values favour open spelling (e.g. the presence of a non-compound-final lexical suffix), others favour hyphenation (e.g. the combination of a lexical and a grammatical constituent) and others again favour solid spelling (e.g. a date of first attestation up to 1500). It is highly likely that most compounds combine a certain number of qualities favouring more than one spelling variant, but in the majority of cases, the combination of these tendencies will result in a clear preference for one candidate. We may conclude, in the wording of Cappelle (2009: 198), that variation in the spelling of individual compounds results "from the opposing influences of

different factors, which thus cancel or subtract from each other, rather than from the absence of any influences". Where the decision for one variant is not clear-cut, language users feel uncertainty and hesitate in their choice. The spelling algorithms, by contrast, have the advantage of always selecting precisely one variant, even if this does not always agree with the spelling preferred by the dictionaries (e.g. for *butter+fly*, which is usually spelled solid but favours open spelling according to Algorithm 4). Depending on the level of perfection aimed at, using the spelling algorithms makes more or less sense. While a predictive accuracy of almost three quarters seems reasonable particularly for less proficient spellers who are uncertain about the most appropriate spelling, more advanced spellers are best advised to use the algorithm as an aid in decision-making merely in cases where their intuition fails them, and where the algorithm represents a considerable improvement on guessing. If the output of the algorithm can be intuitively accepted, it has fulfilled its aim of supporting and simplifying variant selection. If the result of the algorithm cannot be immediately accepted, it means that the person using it has some intuition that they were unaware of; presumably based on the subconscious recognition of some variable's value that only applies to few compounds and is therefore not part of the very general algorithm. For *butterfly*, for instance, the knowledge that idiomatic compounds disfavour open spelling (cf. Table 5.67) would override the open spelling suggested by the pedagogical algorithm. If the most likely spelling suggested by the algorithm contradicts an advanced language user's intuition, selecting the second most likely spelling in the final node of Figure 6.2 is a good strategy (which leads to the selection of solid spelling for *butterfly*). Since even advanced language users usually seem unable to argue why they find certain constructions acceptable or not (Hundt 2010: 40), Table 6.16. summarises the features from Table A.9 which exhibit a strong preference (= 1+; preceded by a black circle) or a simple preference (1; preceded by a white circle) sorted by spelling variant. The features preceded by a horizontal line are statistically non-significant but potentially distinctive for minor groups of compounds. Table 6.16 omits those variables which are included in Algorithm 4, exhibit a strong correlation with the variables retained in the CompSpell algorithm (e.g. a length of two letters correlates with grammatical part of speech) or require the use of reference works or corpora.

Since these exception principles make explicit what advanced language users' intuition may be relying on, they make potentially useful strategies available to less proficient language users and supply arguments to conscious spelling decisions, permitting e.g. to determine the correct spelling

Table 6.16 *Exception principles to the maximally efficient spelling algorithm*

-
-
- Open spelling is preferred
 - if a noun (!) compound is stressed on the second constituent (*black pepper*)
 - if a lexical suffix occurs at the end of the first constituent (*boarding card*)
 - if a compound seems foreign (*anabolic steroid*)
 - if most other compounds ending with the same second constituent are spelled open
 - if the compound is infrequent
 - if the first constituent is an adjective (*main course*)
 - if the compound has a head-final morphological structure (*pot plant*)
 - if the compound is very recent, i.e. dating from 1900 or later (*salad bar*)
 - if the second constituent is capitalised (*girl Friday*)
 - if the compound contains an apostrophe (*banker's order*)
 - if it contains hyphenated constituents, e.g. prefixations or combining forms (*extra-sensory perception*)
 - if one constituent is an acronym (*pay TV*)
 - if the second constituent is a prefixation (*day return*)
 - if one has to type very fast
 - Hyphenation is preferred
 - if the length difference between the constituents exceeds 1:2 measured in letters (*six-shooter*)
 - if there is a very large frequency difference between the constituents (*well-nigh*)
 - if vowels occur at the end of the first and the beginning of the second constituent (*eye-opener*)
 - if the compound ends in the suffixes *-ing*, *-ed* or *-er* (*thought-provoking*)
 - if the first constituent is a grammatical word (*so-called*)
 - if the second constituent is an adjective (*top-heavy*), verb (*spoon-feed*), adverb (*go-ahead*) or grammatical word (*drive-in*)
 - if lexical and grammatical constituents are combined (*far-off*)
 - if the compound is not head-final, i.e. either head-initial (*grown-up*) or unheaded (*know-all*)
 - if most other compounds beginning with the same first constituent are hyphenated
 - if most other compounds ending with the same second constituent are hyphenated
 - if the constituents are identical (*fifty-fifty*)
 - if the first constituent is a verb (*glow-worm*)
 - if the first constituent is inflected (*broken-down*)
 - if a usually open compound is used in attributive position
 - Solid spelling is preferred
 - if the first constituent is stressed
 - if the second constituent is a general reference noun (*weatherman*)
 - if the constituents stand in a species-genus relation (*matchstick*)
 - if the constituents are co-hyponymous (*foxhound*).
-
-

of counter-intuitive spellings suggested by the algorithm. For instance, the feeling that the algorithm's suggested solid spelling for *he+man* is not appropriate can be supported by two arguments: first constituents with two letters (which are usually grammatical words) favour hyphenation, and the combination of lexical and grammatical constituents also favours hyphenation – the unanimous spelling actually found in all dictionaries. The introduction of the exception principles in addition to the maximally efficient Algorithm 4 (or Algorithm 4a for all parts of speech) makes the system more flexible than traditional accounts of English compound spelling.⁶ However, the list presented previously is still too cumbersome to be taught to pupils or students of English in educational contexts, and it was therefore simplified by summarising the most important exception principles for pedagogical purposes (e.g. to be used by teachers).

Note that there are few exception principles to reclassify compounds as solid, since the algorithm misses few of these. In case of doubt, stress should always be used as the first test for open vs. solid spelling. If stress is added to the maximally efficient Algorithm 4, it further splits node 5 of Figure 6.2 and achieves a prediction accuracy of 83.5 per cent with 501 correctly predicted OHS_600 compound spellings (cf. Table 6.18), which is an improvement by 2.8 per cent on Algorithm 4.

However, foreign learners of English (who represent an important part of the target audience) frequently find stress hard to place. That is why stress is suggested as an additional criterion to refine the output, rather than as an obligatory part of the pedagogical spelling algorithm.

For individual compound spelling variants that are not suggested by the algorithm, the successful application of the exception principles suggests that the alternative is acceptable, whereas spellings to which none of the exception principles apply would be classified as inappropriate. This procedure has the advantage that not all variables need to be considered, but only those which tend to predict the spelling that a user prefers over the algorithm's output. The third variant, which is dispreferred both by the algorithm and by the user, can be ignored, unless none of the exception principles should apply to the spelling that the user prefers to the algorithm's output. If more than one exception principle applies to a compound, this makes the alternative spelling

⁶ These work with more specific exceptions and state that adjective compounds with a compound-initial adverb ending in *-ly* favour open spelling, whereas the present account would use the application of the exception principle 'the presence of a non-compound-final lexical suffix favours open spelling' to the output of the algorithm (which predicts hyphenation for adjectives) in order to explain why *politically correct* is spelled open in the data.

Table 6.17 *Simplified exception principles to the maximally efficient spelling algorithm*

<p>Trust your intuition when spelling English compounds. If you find it difficult to decide on a single spelling variant, the spelling algorithm may help you to select the most appropriate one. If the output of the algorithm does not seem right, think of other compounds beginning or ending with the same compound parts and use the most common spelling among those, or check whether the exception principles below support the spelling variant you prefer:</p> <ul style="list-style-type: none">• open spelling<ul style="list-style-type: none">• if a compound is phrase-like (for nouns: STRESS on the second constituent; head-final morphological structure)• if a compound is not lexicalised (literal meaning; low frequency; recent coinage)• if the constituents contain elements disrupting the usual sequence (medial capitalisation, apostrophes, hyphens, suffixes and prefixes at the constituent joint)• if the first constituent is an adjective• if you have to type very fast• BUT NOTE THE BLOCKING PRINCIPLE: even if the exception principles above should apply, adjectives, verbs and adverbs are not spelled open – only nouns and grammatical compounds⁷• hyphenation<ul style="list-style-type: none">• if the constituents are very different (e.g. regarding their length in letters, their frequency or because they combine lexical and grammatical constituents)• if the constituents are identical• if the compound is untypical (not head-final; grammatical words as constituents; compound-initial verb; second constituent is no noun)• if the second constituent ends in the suffixes <i>-ing</i>, <i>-ed</i> or <i>-er</i>• if a usually open compound is used in attributive position (i.e. before a noun)• solid spelling<ul style="list-style-type: none">• if the first constituent is STRESSED• if the second constituent is a general reference noun (like <i>man</i>)• if the first constituent is subordinate to the second constituent (e.g. in <i>matchstick</i>)• if both constituents have the same superordinate (e.g. in <i>foxhound</i>)
--

even likelier. The exception rules favouring open spelling are particularly important, since the algorithm tends to over-predict solid spelling for noun compounds which are better spelled open, e.g. *hot rod*. The more appropriate open spelling is supported by several arguments in this case:

⁷ There are so few adjectives, adverbs and verbs with open spelling (e.g. *politically correct*, *upside down* and *steam clean*) that they can be ignored.

Table 6.18 *Predictive value of Algorithm 4b (including stress) for OHS_600*

		Actual spelling		
		O	H	S
Predicted spelling	O	173	15	31
	H	0	178	19
	S	27	7	150

stress on the second constituent as a major argument for open spelling of nouns, head-final morphological structure and a compound-initial adjective.

An interesting case to consider is the spelling of *sun+lit*, which is incorrectly predicted as hyphenated by the algorithm (because it is an adjective), but which is actually spelled solid in all dictionaries used here. Since the first constituent is stressed, solid spelling is indeed a possible option according to the exception principles. However, some of the principles for open spelling apply too; e.g. literal meaning and head-final morphological structure. In this case, one has to resort to the blocking mechanism listed under the principles for open spelling, which bans open spelling based on the compound's part of speech and introduces an additional degree of accuracy.

To sum up, the spelling algorithm's exception principles can be checked in case of doubt. Stress is the most important criterion overriding all others, but it is in turn overridden by a blocking principle banning open spelling from all parts of speech except nouns. In all other cases, blocking by means of the exception principles is cumulative, in a way that is similar to restrictions in word formation (Bauer 1983: 99):

It seems likely . . . that some of these restrictions (or tendencies) [against which potential formations are checked] may, on occasions, be ignored, but that there comes a point when too many restrictions would be broken if a particular word were formed, and so the potential formation is blocked by a cumulation of factors: that is, the restrictions do not all work independently but in unison when a potential formation is considered. Ideally, it might be possible to speak in terms of the weightings of different restrictions and a threshold level below which restrictions can be ignored.

6.4 Summary

The spelling of a certain number of English compounds can be considered a linguistic case of doubt in the sense of Klein (2003), since it causes uncertainty in the selection of formally similar plausible alternatives and is not easily judged as incorrect in hindsight by competent language users. This chapter considers various heuristics which can be used by spellers who feel uncertain about the spelling of a particular English compound. While guessing has low prediction accuracy, the selection of open spelling as the default with the highest probability of base-line occurrence has the disadvantage of reaching an error rate of 100 per cent for all hyphenated and solid compounds. In order to create more convenient algorithms which model the spelling of English compounds in general, only the statistically significant variables from the empirical study were considered, so as to yield

- **Algorithm 1:** a comprehensive algorithm comprising all statistically significant variables from the training set OHS_600
- **Algorithm 2:** a user-friendly algorithm using only the significant and objective variables which are easy to apply for language users without prior linguistic training
- **Algorithm 3:** a take-the-best algorithm using only the strongest variable, namely part of speech
- **Algorithm 4 (the CompSpell algorithm):** a maximally efficient algorithm aiming for a compromise between large prediction accuracy and a small number of features (plus its variants 4a including grammatical parts of speech and 4b including stress).

The algorithms predict compound spelling based on the majority spelling within the final nodes of their decision trees. The prediction accuracies for all algorithms were tested on OHS_600, the training set. In addition, the maximally efficient Algorithm 4 was tested on the following sets of data:

- **OHS_extra** contains biconstituent compounds with unanimous spelling in all the five or more dictionaries in which they occur (i.e. a test set with particularly strong similarity to the training set)
- **Master_5+_tendency** contains all biconstituent compounds occurring at least five times in the dictionaries with a tendency towards one spelling variant, fewer dictionary occurrences for the second variant and no hits for the third variant

- **CompText** comprises compounds from a corpus of present-day British English compiled specifically for this study which are not contained in the Master List.

The comprehensive Algorithm 1 predicts the spelling of 85.2 per cent of the OHS_600 compounds correctly. The user-friendly Algorithm 2 performs almost equally well with a predictive accuracy of 81.3 per cent, whereas the proportion of compounds spelled correctly by the single-variable Algorithm 3 is very low at 59.2 per cent. By contrast, the predictive accuracy of Algorithm 4, which uses merely three features (part of speech of the compound, length of the compound in syllables and length of the second constituent in letters) is:

- 80.7 per cent for OHS_600
- 76.3 per cent for OHS_extra
- 61.0 per cent for Master_5+_tendency
- 72.9 per cent for CompText.

Algorithm 4 can thus be considered a robust strategy, which remains relatively accurate in the generalization to new data (cf. Czerlinski et al. 1999: 107) and is also simple and easy to apply. When combined with heuristics derived from an additional study of compounded grammatical words, numerals and interjections, the algorithm becomes even more comprehensive and can be expected to predict the spelling of roughly three out of four biconstituent English compounds for all parts of speech correctly. In order to determine how good this prediction accuracy is, the output of Algorithm 4 was compared to the spellings favoured by two native speakers of English with a background in language teaching. The results were so similar that the CompSpell algorithm can be considered a very good approximation to native speaker competence for English compound spelling. Due to its simplicity and efficiency, it lends itself very well to the teaching of English as a foreign language and has advantages over and above the use of dictionaries, grammar books and style guides for the spelling of English compounds. However, since the algorithm predicts about a quarter of the dictionary spellings incorrectly and may thus also confuse potential users tending towards a specific (possibly even more appropriate) spelling, the CompSpell algorithm is only intended as a supplementary strategy for the spelling of compounds for which no conclusion can be arrived at by intuition alone. Since the intuition particularly of more advanced spellers may be based on more variables than those retained in the pedagogical algorithm, this chapter also

introduces a number of exception principles to support conscious spelling decisions (of which stress is the most important criterion overriding all others) and a blocking principle which bans open spelling from all parts of speech except nouns, and which overrides the exception principles in turn. Since some spellings may be instances of free variation, even a number of seemingly incorrect predictions by the algorithm will presumably be acceptable to many language users.

Modelling English Compound Spelling

Chapter 6 provided a statistical treatment of English compound spelling with a focus on heuristics for pedagogical application. This is complemented in the following sections by a more theoretical treatment of the phenomenon, which considers how existing models of language can be applied to the spelling of English compounds in the light of the evidence from the empirical study. This chapter discusses the suitability of prototype-based, analogical and cognitive approaches for the modelling of the present state of English compound spelling and closes by considering how linguistic change can be integrated into such models.

7.1 The Relation between the Three Main Spelling Variants

A very basic issue to consider in the modelling of present-day English compound spelling is in what relation the three main spelling variants stand to each other. This, in turn, requires the consideration of whether the individual compound spelling variants carry any meaning. There is a common – although not universally accepted – assumption that a difference in form can be regarded as the result of a difference in meaning or function (cf. e.g. Goldberg’s 1995: 67 principle of no synonymy). One may therefore wonder whether the three major compound spelling types encode any particular information distinguishing them and justifying the existence of so many variants. An attempt to answer these questions is made in the following.

As we have seen, the three major spelling variants are distinguished by the use or non-use of the hyphen or the space (cf. 2.5.4). Huddleston and Pullum’s (2002: 1724–1726) system of punctuation marks as segmental units which “occupy a position in the linear sequence of written symbols” and “give indications of the grammatical structure and/or meaning of stretches of written text” includes hyphens and spaces (which only become visible by means of the segments surrounding them; cf. Fries 2012:

Table 7.1 *The graphical realisation of punctuation indicators (following Huddleston and Pullum 2002: 1725–1726)*

Indicator	Example	Realisations
(Ordinary) hyphen	<i>non-negotiable</i>	hyphen character (not flanked by spaces)
Long hyphen	<i>the doctor–patient relationship</i>	en rule (not flanked by spaces)
Dash	<i>He’s late - he always is.</i>	hyphen character (flanked by spaces)
	<i>He’s late – he always is.</i>	en rule (flanked by spaces)
	<i>He’s late—he always is.</i>	em rule (not flanked by spaces)

401–414), but not concatenation. The graphical shape of the punctuation indicators may vary (Huddleston and Pullum 2002: 1724–1726), with the hyphen character occasionally taking over the functions of the long hyphens but not vice versa (cf. Table 7.1): in Huddleston and Pullum’s (2002: 1729–1730) model, the punctuation indicators have several main functions, the first of which is considered primary:

- **indicating boundaries**, i.e. making clear where particular entities end and where others begin
- **showing status**, e.g. when a question mark specifies a sentence as a question
- **signalling omission**, e.g. by means of an asterisk or an apostrophe in *f*ck* or *won’t*
- **indicating linkage**, e.g. by means of hyphens
- **the prevention of misreading**, e.g. by adding a comma to the sentence *Liz recognised the man who entered the room (,) and gasped.*

When considering with what functions the punctuation marks are used in the three major compound spelling variants, the space mainly seems to assume the boundary-indicating function. Concatenation, at the other extreme, can only signal unity status due to its maximally linked state. The hyphen, by contrast, has a multitude of recognisable functions,¹ of which the following are relevant for compounds:

¹ Hyphens may e.g. indicate the prolongation of vowels (*He-e-elp!*; Quirk et al. 1985: 1636), stammering (*P-p-p-please t-t-try!*; Quirk et al. 1985: 1636), exaggeratedly slow and careful pronunciation (*Speak c-l-e-a-r-l-y!*; Huddleston and Pullum 2002: 1759), syllabification (*punc-tu-a-tion*; McDermott 1990: 115) or the spelling out of a word letter by letter (*l-i-a-i-s-o-n*; Merriam-Webster 2001: 35).

- Hyphens indicate the boundaries of compound constituents (which is reflected in their treatment under the subheading of “Separation” in Quirk et al.’s 1985 Appendix III, “Punctuation”).
- Since a non-line-final hyphen indicates that a construction is no simple word or phrase (cf. Carney 1994: 48), it can be regarded as marking word formation status at least to a certain extent.
- Hyphens may indicate omission (more specifically, that of the second constituent) in elliptical compounds such as *car-* and *ship-owners* (cf. 2.5.4).
- The hyphen’s function of indicating linkage is present in all instances of English compound spelling and also in end-of-line hyphenation. This function is so important that it is even maintained across a blank in instances where the hyphen signals omission, e.g. in elliptical compounds.
- In style guides, the use of hyphens is frequently advocated to prevent misreading, as in Hanks’ (1988) example *a machine-tool minder* vs. *a machine tool-minder*, or when the hyphen makes all the difference between compounds with an initial *-ing* form (*There is a **moving-van*** = large car) and otherwise identical phrases (*There is a **moving van*** = a van in motion; cf. Bailey 1979: 4).

In addition to the coverage of all of Huddleston and Pullum’s functions, one may recognise one more function that the hyphen assumes with respect to compounds, namely the indication of **markedness**. Markedness may refer e.g. to an unusual morphological structure, the combination of heterogeneous elements, the novelty or ad-hocness of the compound (cf. Fowler 1921: 5; Carney 1994: 48; Morton Ball 1939: 22) etc. (cf. 5.12.7). The observation that hyphenation is the most marked of the three compound spelling variants can be explained by the fact that hyphenations contain additional physical material compared to the other two types (just as the more complex plural form in English is marked relative to the simpler and more basic singular form containing fewer morphemes; cf. also Croft 2002: 89).

The markedness of hyphenated compounds seems to attract a particularly large amount of criticism: the advice that unnecessary hyphens should be avoided can be found frequently in style guides (e.g. Fowler 1921: 5; Morton Ball 1951: 3; Cullen 1999: 51) in spite of the fact that this statement is tautological, because it contains the evaluative word *unnecessary*: regardless of the precise amount of hyphenation favoured, most people would presumably agree that once a certain level is exceeded,

practically anything can become detrimental. While Morton Ball (1951: 8) argues that “[a] hyphen should be used in compounding only to facilitate understanding or to denote temporary expediency”, other authors (e.g. Hanks 1988) have gone so far as to advocate the complete abolition of the hyphen in compounds. The unfavourable treatment of hyphenated compounds in the literature may be linked to their intermediate status “between separation and fusion” (Fowler 1921: 8), which might call more strongly for a clarification in either of the two more extreme directions, and possibly also to the fact that a critical view of hyphenation is already expressed by Fowler (1921: 5) as one of the most important prescriptive writers on the English language. Furthermore, patterns with high type frequency are judged as more acceptable than patterns with low type frequency (Bybee 2003: 13), and hyphenations occurred least frequently in the empirical study.

When contrasting the three major English compound spelling variants with the aim of determining the relation between them, the fact that the opposition is not binary but ternary suggests the consideration of each spelling variant in contrast to the group combining the other two, so as to reveal the similarities and differences between them. Tying in with this, open and solid spelling contrast with hyphenation by the fact that they seem to be generally regarded as unmarked spellings. Open and hyphenated spelling both reveal and result in a stronger awareness of the compounds’ constituents (viewed from a productive and receptive perspective, respectively) than solid spelling. Solid and hyphenated spelling, by contrast, share that the unit status of compounds with these spellings is unquestioned, whereas open spelling occasionally allows for reinterpretation as a phrase (cf. 2.1). To conclude, three different subgroups can be established based on the following parameters:

- a) unmarkedness (open and solid spelling)
- b) analyticalness (open and hyphenated spelling)
- c) unit status (solid and hyphenated spelling).

This can be illustrated graphically, as in Figure 7.1.

Since some compounds may occur with more than one spelling, from a logical perspective, seven distribution patterns are possible for individual English compound types (cf. Table 7.2).

While Sepp (2006: 10) uses the term *free variation* where two or more spellings are being used, it is worth exploring whether this can also be

Table 7.2 Sepp's (2006: 86) seven logical distribution patterns for English compound spelling

Open	Hyphenated	Closed
+	–	–
–	+	–
–	–	+
+	+	–
+	–	+
–	+	+
+	+	+

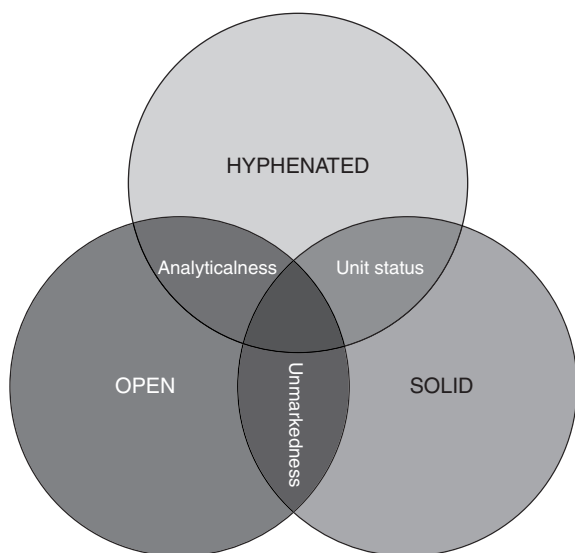


Figure 7.1 The relationship between open, hyphenated and solid spelling

considered an instance of *allography*: graphemes are “the smallest units in a writing system capable of causing a contrast in meaning” (Crystal 1997: 196), and their formally variant realisations are known as *allographs*. However, as far as the form side of the ‘units’ under consideration is concerned, only hyphens and spaces have physical extension, whereas solid spelling has no orthographic indicator of compound status, even though

a comparable meaning/function ‘compound constituent juncture’ is involved.² By contrast, all three spelling variants fulfil the other requirement of Crystal’s (1997: 196) *grapheme* definition, that of causing a contrast in meaning. Various dictionaries, style guides and grammars provide examples for spelling-induced minimal pairs:

- *air line* ‘pipe supplying air’ vs. *airline* ‘air transport company’ (Waite 1995)
- *half tone* ‘(AE) semitone’ vs. *halftone* ‘a way of printing black-and-white photographs that uses spots of different sizes’ (MED)
- *hard-core* ‘detailed and often violent pornography’ vs. *hardcore* ‘electronic music that is loud and fast and sounds aggressive’ (MED)
- *in box* ‘(AE) *in tray*’ vs. *inbox* ‘the place on a computer program where emails arrive’ (MED)
- *launch pad* ‘a base from which a weapon or spacecraft is sent up into the sky’ vs. *launchpad* ‘a place on the internet that helps you start to find information about a particular subject’ (LDOCE)
- *sleeping partner* ‘inactive business associate’ vs. *sleeping-partner* ‘that you sleep with’ (Carney 1994: 49)
- *slip case* ‘(AE) plastic container used for a CD or DVD’ vs. *slipcase* ‘a hard cover, like a box, for putting a book in’ (LDOCE).

By analogy to the traditional distinction between phones and phonemes, a single meaning-distinguishing context could be deemed sufficient to postulate graphemic status (but cf. Herbst’s 2010: 60 criticism regarding the status of /θ/ vs. /ð/). Since the foregoing examples cover all possible oppositions (open vs. hyphenated, solid vs. hyphenated and open vs. solid), the space, the hyphen and concatenation could therefore be treated as different graphemes (even though the last of these is atypical in having no physical extension). At the ends of lines, however, only open spelling is distinguishable from the other two types, and the distinction between solid spelling and hyphenation is neutralised, since the horizontal line may realise either a hard or a soft hyphen (by analogy to the neutralisation of phonemes in phonology, e.g. when all syllable-final obstruents in German are automatically devoiced and strengthened; cf. Eckert and Barry 2005: 43–44).

² Note that there does not seem to be an established cover term for the entities occurring or absent at the constituent joint: *connector* is inappropriate, because open spelling has no connecting function, and *separator* is inappropriate, because solid spelling does not orthographically mark the constituent boundaries.

However, one may wonder how frequently successful communication is really at stake due to the choice of an unusual or incorrect compound spelling variant. In general, it seems that few language users make compound meaning depend on spelling alone, and Hanks (1988) finds no instances of this in the Cobuild corpus. Nevertheless, some real, unconstructed misunderstandings are also reported (e.g. by Truss 2003: 171–172), and Goldstein (2004: 323) points out that the hyphen makes all the difference between words of “dual heritage”, such as *Mexican-American*, and determinative compounds such as *Latin American*. A *Mexican American* with open spelling may therefore be assumed to feel more like an American with Mexican roots, whereas a *Mexican-American* will presumably place both origins on the same level of importance. To conclude, the use of the three types of spelling can lead to meaning differences – even though only in a clear minority of cases.

7.2 **Prototype Theory**

Prototype theory has already been applied to different areas of linguistics (e.g. past tense, causation and dative alternation; cf. Taylor 1989, Gilquin 2006 and Gries 2003b), and its potential for the modelling of English compound spelling is explored in the following. On a general level, prototype theory posits that exemplars within categories are clustered around a central prototype, whereas the categories’ periphery is characterised by fuzzy boundaries (cf. the seminal work by Rosch 1973, 1975). How precisely the prototype should be defined is, however, a matter of dispute. Gilquin (2006: 180) concludes from her critical discussion that the prototype is best described as

a multi-faceted concept, bringing together (1) theoretical constructs found in the cognitive literature and relying on deeply-rooted neurological principles such as the primacy of the concrete over the abstract, (2) frequently occurring patterns of (authentic) linguistic usage, as evidenced in corpus data, (3) first-come-to-mind manifestations of abstract thought, as revealed through elicitation tests and (4) possibly other aspects that contribute to the cognitive salience of a prototype

– by which one might understand e.g. productivity, transparency and naturalness (cf. Winters 1990: 291–296). Even though frequency is not identical with prototypicality, prototypicality is commonly operationalised in terms of frequency of occurrence in linguistic studies (cf. Gilquin 2006: 168–169).

Prototype theory can be applied to English compound spelling from various perspectives, e.g. by considering which spelling variant is the most prototypical one or, alternatively, what constitutes a prototypical compound and what spelling is preferred by that subgroup of compounds (cf. later in this chapter). The notion of prototypicality permeates the whole empirical study presented here: in contrast to previous research on the spelling of English compounds, which focuses on the reasons for variation and thus the periphery of the categories of open, hyphenated and solid spelling, the present study's approach is based on the assumption that there are compounds with a prototypical spelling (which can be determined by their coincidence in various dictionaries). It is assumed that the compounds belonging to each of the three groups (exclusively open, exclusively hyphenated and exclusively solid spelling) share certain group-internal characteristics (i.e. the values of the features coded as variables in the database), and that these characteristics can be used to establish prototypes for open, hyphenated and solid compound spelling, respectively. The peripheries of the three spelling variants are fuzzy, since some compounds in the English language may occur with open or solid spelling, others with open or hyphenated spelling, yet others with hyphenated or solid spelling and some even with all three variants. Depending on the differences in the frequency of occurrence of actual usage tokens for each of the three variants, individual English compounds are situated more or less closely to the centres of the three spelling variant categories.

The compounds in Table 7.3 are used as examples in the discussion of how the spelling of English compounds can best be modelled in a prototype account. The capitalisation in the first column indicates which spelling dominates across the reference works; the precise frequency for each variant can be found in the last three columns. While the following models use dictionary-based frequencies, the approach can also be applied to corpus data.

Since Rosch's original groundbreaking articles on prototypicality do not use any emblematic graphical representations, the first attempt at modelling prototype structure for a compound spelling variant in Figure 7.2 draws heavily on the presumably most famous diagram exemplifying prototype structure, namely Aitchison's (1994: 54) concentric circles representing the category BIRD.

This representation of the prototype structure of the category ENGLISH COMPOUNDS WITH SOLID SPELLING is supposed to show that *firewall* is a highly prototypical compound: with six solid as against no open or hyphenated spellings, it occupies the most central position

Table 7.3 *Selected OHS_600 and Master_5+ compounds with spelling-sensitive frequencies across dictionaries*

Spelling frequency pattern	Example	O	H	S
OHS	<i>front+runner</i>	2	2	2
OHs	<i>brain+teaser</i>	2	2	1
OhS	<i>whole+wheat</i>	2	1	2
oHS	<i>zoo+keeper</i>	1	2	2
OH	<i>can+opener</i>	3	3	0
Oh	<i>egg+timer</i>	5	1	0
oH	<i>well+paid</i>	2	4	0
OS	<i>lawn+mower</i>	3	0	3
Os	<i>guest+room</i>	4	0	1
oS	<i>summer+house</i>	2	0	3
HS	<i>best+seller</i>	0	3	3
Hs	<i>night+time</i>	0	4	2
hS	<i>news+stand</i>	0	1	5
O	<i>elastic+band</i>	6	0	0
H	<i>by+product</i>	0	6	0
S	<i>fire+wall</i>	0	0	6
S	<i>fire+fly</i>	0	0	5

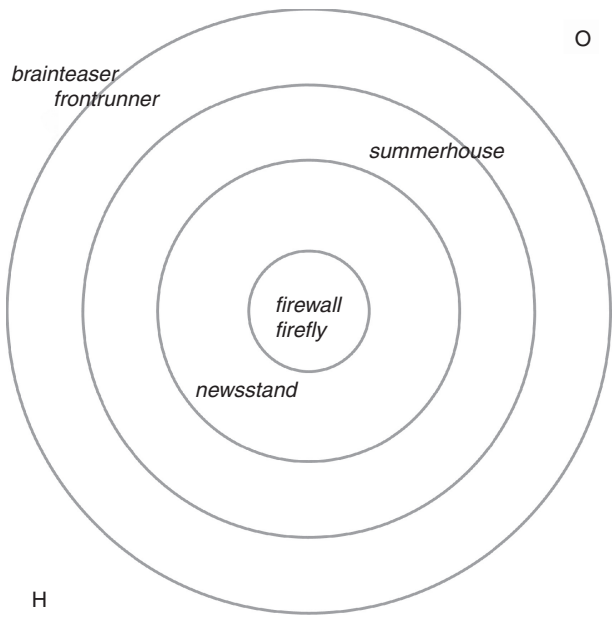


Figure 7.2 Preliminary prototype structure of the category ENGLISH COMPOUNDS WITH SOLID SPELLING

among the compounds in the example dataset (cf. Table 7.3). Since the compound *firefly* also occurs exclusively with solid spelling, but only five instead of six times, this raises the question whether it deserves to be considered equally prototypical or slightly less so. In order to take absolute frequency into account (in addition to cross-categorical proportional frequency) while still noting the absence of occurrences with alternative spellings, *firefly* is classified as a category-central compound (exemplified by its position in the innermost circle of Figure 7.2), but slightly removed from the absolute centre by comparison to *firewall*. With five solid spellings and one hyphenated occurrence, *newsstand* is a more prototypical solid compound than *summerhouse* with three solid and two open spellings – which is represented by *newsstand*’s more central position in the concentric circles. Since the two compounds also differ with regard to their spelling alternatives, *newsstand*’s slight inclination towards hyphenation is additionally indicated by its placement on a diagonal on which increasing distance from the category centre of solid spelling indicates an increasing tendency towards hyphenation (cf. Figure 7.3). The same system is used to visualise *summerhouse*’s slight tendency towards open spelling.

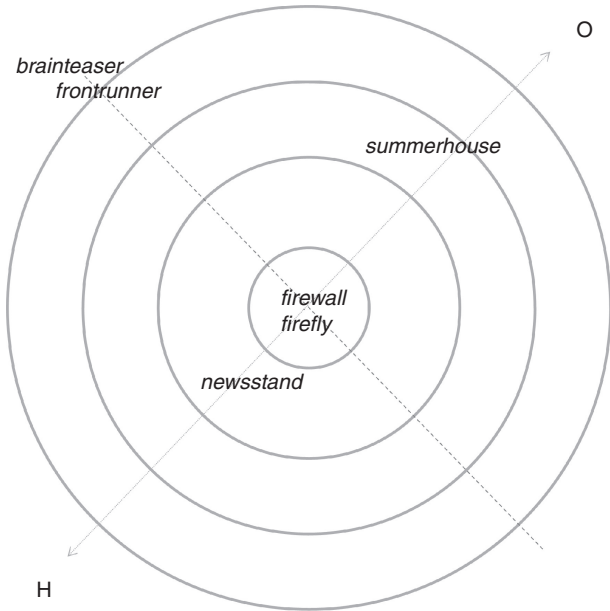


Figure 7.3 Modified prototype structure of the category ENGLISH COMPOUNDS WITH SOLID SPELLING

This modification of Aitchison's (1994: 54) original diagram is made necessary by the fact that traditional prototype accounts merely consider the phenomenon from the perspective of membership of a single category: either an exemplar belongs in the category under consideration or not – or it is situated on the category's fuzzy boundaries. The application of prototype theory to the spelling of English compounds, by contrast, does not merely involve the classification as 'solid' (in the previous example), 'not solid' or on a gradient between the two, but additionally requires the consideration of the alternative spelling options which are present for language users attempting a classification as open, hyphenated or solid in the writing process. That distinguishes COMPOUND SPELLING with its three subcategories OPEN SPELLING, HYPHENATED SPELLING and SOLID SPELLING from other categories (such as BIRD), in which the alternatives may exhibit more variation: thus some birds might contrast with bats; kolibris might contrast with insects (because they are so small and move in such a special way) and penguins might seem fish-like in certain respects. When a new, unknown animal is considered by a language user, the animal kingdom offers so many options for classification that the opposing alternatives are not always the same for each animal. In the spelling of English compounds, by contrast, there are usually only three options – and one of them has to be selected in the spelling process in spite of possible multiple category membership. The graphical representation used in Figure 7.3 also has the advantage of permitting the depiction of *front+runner* as a peripheral solid compound: since it occurs with equal frequency in all three variants, it is situated on the outer circle of the category of solid spelling, on the dotted line indicating equal distance to the centres of the categories OPEN SPELLING and HYPHENATED SPELLING. *Brain+teaser* is even more peripheral in the sense that open and hyphenated spelling are equally likely and more frequent than solid spelling. This is represented by the fact that the orthographic representation of the compound merely touches the outer circle of the category SOLID SPELLING in Figure 7.3. However, this depiction also has a number of disadvantages: on the one hand, the bottom right corner (i.e. where the dotted line ends) remains empty. Of course, it would be possible to fill the space with the types of compound placed at the top left corner so far, by using the arrow O-H as the axis of symmetry, particularly when charting larger numbers of compounds. However, that would be equally arbitrary – which suggests that the diagram in its current form is not ideal. More importantly, the current visualisation only focuses on solid spelling category membership. As a consequence, *brain+teaser*, which fluctuates more between hyphenated and open spelling (with two

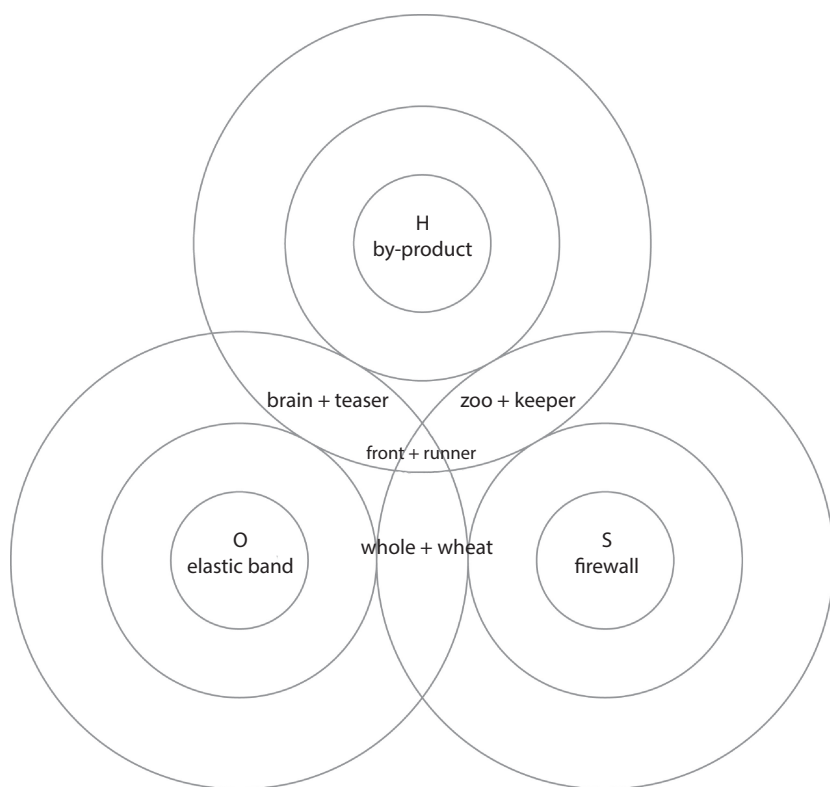


Figure 7.4 Overlapping prototype model for compounds with open, hyphenated and solid spelling

hits each) and only has a single solid attestation in the data, can merely be depicted as peripheral regarding solid spelling, but the diagram does not really show its degree of simultaneous inclusion in the other two categories. All this seems to call for a different format. If we keep the idea of concentric circles, the most plausible solution consists of three overlapping categories with areas of different strength. As before, the centre of each circle represents the centre of the category (cf. Figure 7.4).

In this model, the compound *front+runner* is situated in the middle of the region where all three categories overlap. The compounds *zoo+keeper*, *whole+wheat* and *brain+teaser*, which occur in all three spellings but vary equally between the two most frequent ones, are moved to those regions within that central area which are most distant from open, hyphenated and

solid spelling, respectively. By contrast, *guest+room* would be situated as centrally as possible in the category OPEN SPELLING, while permitting minimal overlap (for a frequency of one) with SOLID SPELLING, in order to indicate variation between merely two spellings. By analogy, one could enter the precise position of all biconstituent compounds, not only from the empirical study but (provided that usage frequency data are available) in general. This would result in a comprehensible and easy-to-interpret diagram. However, this graphical representation has a very important flaw as well: the regions at the top, bottom left and bottom right would remain empty, because the highest level of category membership lies in the middle of each concentric circle and because exemplars with decreasing category membership are situated closer to the contrasting categories. Upon closer consideration, it becomes clear that the relationship between the contrasting categories would actually require the highest level of category membership to be as distant from the other two categories as possible. As a consequence, it should not be situated in the middle of the circles but rather in the directions in which the arrows in Figure 7.5 are pointing.

As soon as this major change in perspective has been implemented, we get an extremely flexible prototype model: we can now leave the two-dimensional plane and enter the third dimension by reinterpreting the arrows in Figure 7.5 as the axes of a three-dimensional coordinate system (cf. Figure 7.6).

In such a system, each point – and thus the position of each compound – is specified by three coordinates, namely the frequencies of occurrence with open, hyphenated and solid spelling, respectively (with the assignment of spelling variant to the x-, y- and z-axes being arbitrary). Since *news+stand* occurs zero times with open spelling, one time with hyphenated spelling and five times with solid spelling, its position (0, 1, 5) corresponds to a precisely determined point on the plane formed by the axes ‘Hyphenated’ and ‘Solid’. If its position is charted by beginning at the point where all three arrows meet, one has to move zero units in the direction of the ‘Open’ arrow, one unit to the top, in the direction of ‘Hyphenated’, and five units to the right, in the direction of ‘Solid’. The same procedure can be followed for any other compound. This way of modelling compound spelling, which is based on spelling-sensitive absolute frequency, has several consequences and implications: for instance, the best exemplar is relegated to infinity. All compounds situated on the arrows are central members of the category represented by the arrow, since they have a frequency of zero for the other spelling variants. At the same time, the model recognises an internal gradient based on their frequency (e.g. *firewall* with a count of six as against *firefly* with a count of

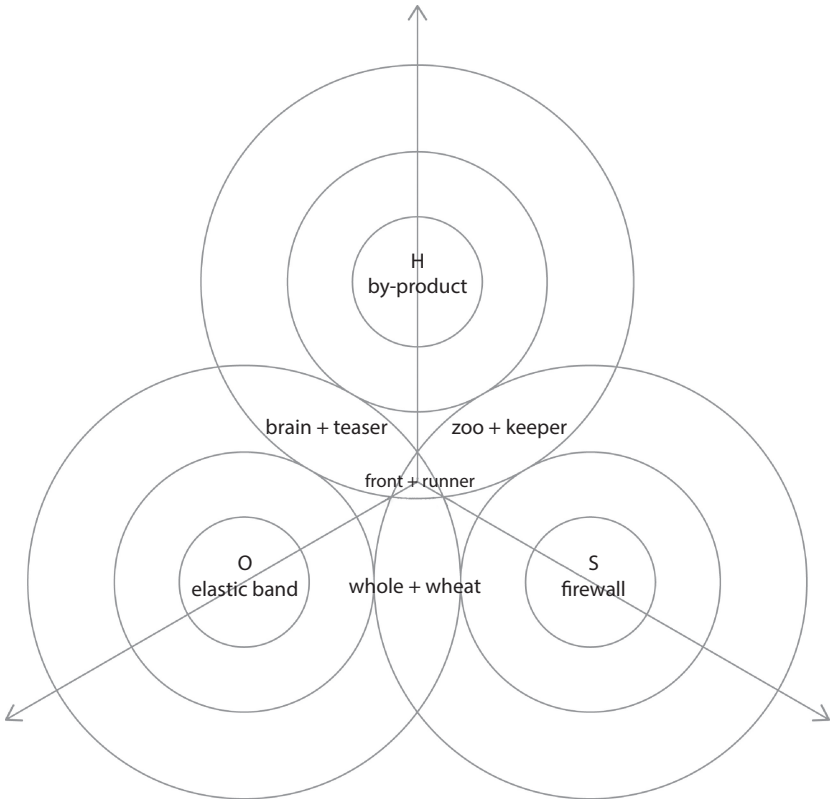


Figure 7.5 Modified overlapping prototype model

five), which leaves room for the possible existence of even more prototypical exemplars (in terms of an even more extreme frequency difference between the three categories) without having to question the whole model, and it simultaneously represents the idealised nature of the prototype (cf. e.g. Gries 2003b: 22). The point where the three arrows meet represents absolute zero, i.e. compounds using none of the three major spelling variants (because they employ a slash or other alternative spellings; cf. 2.5.4). All compounds which occur equally frequently with all three spellings are situated on a line whose endpoint is absolute zero, but this cannot be rendered graphically in Figure 7.6, since the line with the value $O = H = S$ runs perpendicularly into the view of the observer and is therefore indistinguishable from the point at the bottom. However, this could be easily remedied by placing the

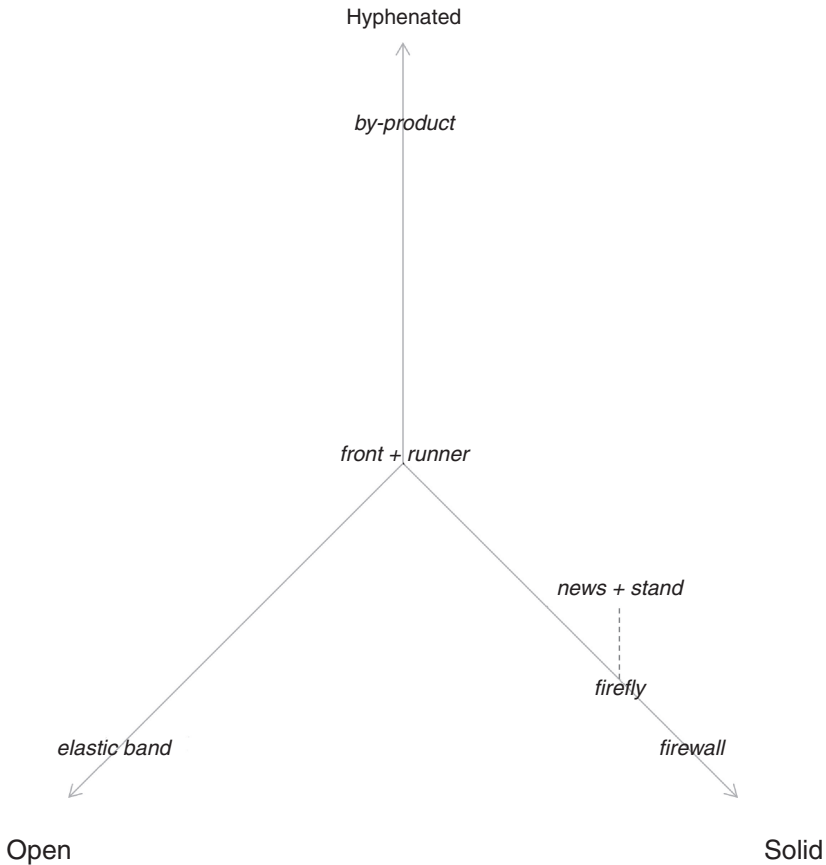


Figure 7.6 A three-dimensional prototype model of English compound spelling

three axes at a slightly different angle in a modified diagram. The other three axes of doubt with equal likelihood of two spelling variants also run through point zero and are equally distant from two of the variant arrows each. If more than the few examples from Table 7.3 are used, it is possible to create an impressive three-dimensional scatterplot. The program R even permits the creation of interactive three-dimensional scatterplots which can be rotated with a mouse by using the `plot3D(x, y, z)` function in the `rgl` package or the `scatter3d(x, y, z)` function in the `Rcmdr` package (www.statmethods.net/graphs/scatterplot.html, 07 November 2013).

For representational convenience, frequency of usage of individual compounds with particular spellings has been used as the sole variable so far. The resulting model places compounds with similar behaviour in a structured spatial relation to each other and permits the fast identification of the compound types which are most characteristic of each of the three spelling variants and the most common spelling variant as the spatial area with the largest number of compounds. While this model is still limited in its representative power, the large number of variables exerting some influence on the spelling of English compounds can be integrated very easily by considering that each compound (and thus each data point in space) represents an array consisting of key-value pairs consisting of the variables (such as word length, part of speech etc.) as the keys and the corresponding values. Thus the compound *firefly* comes with information on word length (three syllables), part of speech (noun), stress (on the first constituent) etc. In a formal computational representation, this information could be coded systematically, i.e. in the same order, for all known variables (3, n, 1, . . .). By analogy to some electronic dictionaries (such as the OED) which allow the user to select the type of information (e.g. pronunciation or etymology) to be shown, it should be possible to focus on a particular aspect (such as stress) to be shown for all compound data points. If the various values or numerical ranges for each variable are coded with different colours (e.g. stress on the first constituent in red and stress on the second constituent in green), this should result in a very intuitively understandable three-dimensional model which conveys visually what characterises each of the three spelling variants. Most importantly, this approach yields easily accessible information on what proportion of data points with a certain spelling have a particular value, how these are situated in relation to each other and thus e.g. on whether a certain value is more common among the compounds on (or close to) a particular variant axis or among those compounds with alternative spellings.

The traditional prototype model considers the degree of membership for one particular category only. Even if the famous bird diagram uses a circular plane, it actually expresses a one-dimensional relation which could be depicted equally well by using a simple linear scale such as the individual arrows in Figure 7.6. In the present model, membership with relation to two contrasting categories is expressed by using a plane (which corresponds to a two-dimensional coordinate system with an x- and a y-axis). The three-dimensional model of compound spelling was outlined earlier, and from there, it is only a small step to a multidimensional prototype model: depending on the requirements and the number of categories to be

considered in relation to each other, one may construct a coordinate system with any number of axes. We simply need as many coordinate values for each exemplar as there are axes – e.g. seven values if seven categories are compared. How useful this is for the modelling of compound spelling may at first be called into question, since additional spelling alternatives are quasi non-existent (cf. 2.5.4), but a multidimensional model is necessary to chart possible spellings for compounds with more than two constituents, since triconstituent compounds have nine possible spellings (OO, HH, SS, OH, OS, HO, HS, SO, SH), and compounds with four constituents are already characterised by twenty-seven possible spelling combinations (e.g. open+hyphenated+open).

While the application of prototype theory to compound spelling discussed earlier involves a considerable modification of the usual representation of prototypes, its application to other aspects, e.g. what can be considered the most prototypical of the three major spelling variants, takes more traditional forms. Based on the results of the empirical study, and using frequency as an operationalisation, it is possible to conclude that open spelling is most frequent among the compounds recorded in dictionaries (and their occurrence in corpora) and may act as a kind of prototype in this respect.

Furthermore, one may attempt to determine what spelling variant is favoured by the most prototypical English compounds. While some strict definitions of prototypicality maintain that the prototype should possess all the prototypical features (e.g. Cruse 1990: 291), other, looser definitions (like that by Givón 1986: 79) only expect it to possess “the greatest number of *important* characteristic properties”. In the empirical study, we have considered a large number of possible variables (and their respective values) which may play a role in the spelling of English compounds, but some of these seem more important than others: if we are to judge from the amount of research devoted to nominal compounds consisting of noun+noun, this feature should certainly be included in the prototype. Keeping in mind the literature discussed previously, it also seems safe to claim that prototypical English biconstituent compounds are lexicalised, head-final and stressed on the first constituent. If the OHS_600 compounds are filtered according to that combination of criteria, we get a list of 244 compounds, of which 113 (46.3 per cent) are spelled open, 5 (2.0 per cent) are hyphenated and 126 (51.6 per cent) are spelled solid. From this, we can conclude that the prototypical English compound is definitely not hyphenated. The results for open and solid spelling, by contrast, are so close that it is difficult to concede the status of the sole prototype for English compound spelling to

solid spelling. More likely than not, we are dealing with a twofold prototype – and that is precisely why language users find it so difficult to spell English compounds. While solid spelling may represent a kind of prototype based on explicit teaching (e.g. when many linguistics textbooks discuss the distinction between the solid compound *blackbird* and the open phrase *black bird*), open spellings are far more frequently encountered in language as a whole (cf. Chapter 6). From a cognitive perspective, solid spelling may suggest indeed that the speller considers the compound a unified entity, but open spelling should not be understood as necessarily indicating the opposite (in spite of its usual boundary-delimiting function), because some variables like increased length may prevent the solid spelling even of idiomatic compounds such as *glove+compartment* (cf. 5.8.2.3). All of this seems to suggest that traditional accounts of prototypicality reach their limits in the modelling of English compound spelling. Since we have seen in the empirical study that compound spelling variant selection follows probabilistic rather than deterministic principles, the whole system is also reminiscent of Wittgenstein's (1972: 31–32) family resemblances (cf. the discussion in 5.12.7), with not all category members of open, hyphenated or solid spelling necessarily sharing one particular feature/value combination but rather exhibiting a network of overlapping qualities.

7.3 Analogy

Analogy, “the process by which novel utterances are created based on similarity to previously experienced ones” (Rovai 2012: 188), is another concept that is frequently used to explain linguistic behaviour. Analogy relies on memory rather than on abstract analyses (Bybee 1998: 433) and works in two ways: either the spelling of a ‘new’ (i.e. low-frequency) compound is based on the spelling of one particular high-frequency compound (which can be considered a prototype for all compounds following its spelling) or the spelling of the target compound is modelled on a pattern which emerges from the general behaviour of compounds that are related to the word in question, e.g. in terms of their components or their meaning (which can be considered a governing principle). In this account, “[g]eneralizations over forms are not separate from the stored representation of forms but emerge directly from them” (Bybee 2003: 7). Bybee (2003: 52) argues that individual tokens of language use necessarily have to be stored in the mental lexicon at least for a certain period so as to permit categorisation. Since recurring patterns are more prominent than

infrequent ones, they are expected to be more productive (Bybee 2003: 6–7). Support for this comes from the results of the empirical study, in which open spelling was used not only for the majority of the 9,258 biconstituent Master List compound types (of which 5,118 occur exclusively with open spelling, and 1,100 more are spelled open in at least one dictionary) but also dominated in the innovative, unlisted compounds from the CompText corpus (63.8 per cent).

The exemplar- and frequency-based principle of analogy, which blurs “the traditional dividing line between grammar and language processing” (Plag, Kunter and Lappe 2007: 205), is often discussed in opposition to rule-based approaches (in which the term *rule* does not necessarily have any prescriptive implications but may be understood in the sense of underlying patterns or governing principles). One very important advantage of rules lies in their economical application, since a single rule may be used to determine the form of many linguistic items (cf. Kreiner and Gough 1990: 104). This would seem of particular importance for beginning learners of a language, who can only draw on a small set of stored material for analogical comparisons. However, the irregularity of the English language presumably prevents its exhaustive description by generally applicable principles (cf. also Kreiner and Gough 1990: 105–106).

Analogical accounts of language have been shown to outperform rules in a number of studies. For instance, Plag, Kunter and Lappe (2007), who investigate stress placement in noun+noun compounds and use the memory-based learning system TiMBL 5.1 (Daelemans et al. 1997; for the most recent version, cf. <http://ilk.uvt.nl/timbl/>), assign stress based on the stress pattern that was most frequently found among the nearest neighbours as the most similar forms (Plag et al. 2007: 222). They compare their own results with rule-based models and the findings of several recent studies on compounding, concluding that “[p]robabilistic and analogical models have higher overall accuracy rates” and that “variable compound behavior is best accounted for by probabilistic or analogical models, instead of rule-based ones” (Plag et al. 2007: 227). Similarly, Krott, Baayen and Schreuder’s (2001: 69) research on linking elements in Dutch finds “unambiguous evidence for a strong analogical effect of the constituent family” and claims to provide a “computational model of analogy with a higher degree of observational adequacy than can be achieved by means of the rules proposed in the literature” (Krott et al. 2001: 52). However, if we assume that the form of new compounds is based on that of stored exemplars, this has one very important implication: the search for similar exemplars requires a relatively uneconomical scan of the whole mental

lexicon (Krott et al. 2001: 76). The more variables analogical language production is based on, the more important this drawback becomes. As far as the required computing power in analogical simulations is concerned, Skousen (2002: 45) points out that “increasing the given context by one variable doubles the memory requirements as well as the running time”. In view of the present study’s observation that many variables may play a role in the determination of variant selection, the calculation of distance metrics for all items in the mental lexicon seems unlikely from a psycholinguistic perspective (Krott et al. 2001: 80–81), because “human decision-makers have limited ability to attend to multiple cues and . . . process information sequentially rather than integratively”, with simple heuristics leading to particularly good decisions compared to regression-based models “under conditions of uncertainty or limited information” (Green and Mehr 1997: 224). For that reason, Krott and colleagues (2001) use TiMBL (cf. earlier in this chapter), which follows a method that is similar to that of the CompSpell algorithm, and also construct a decision tree from a training set. Since a decision tree constitutes a set of rules (Krott et al. 2001: 81), this blurs the dividing line between analogical and rule-based approaches.

The impact of analogy on English compound spelling can be determined by examining the results for the analogical variables that were investigated in the empirical study (cf. 5.10.3). While left and right constituent family size and left and right constituent family frequency reach the level of significance in the statistical analyses, their importance as determinants of spelling variant selection should not be overestimated (cf. 6.2): the comprehensive algorithm, which automatically filtered out correlated variables, only retains left and right constituent family size (cf. Figure 6.1), both of which occur in the final node of the decision tree and only affect a small number of compounds: right constituent family size affects twenty-seven compounds, and left constituent family size does not even affect the majority decision in a group of fifteen compounds. As a consequence, the maximally efficient algorithm does not retain any analogical variables, since these correlate with other variables exerting a stronger influence on compound spelling variant selection.

The results for the analogical variables are particularly striking when compared to those for the rule-based algorithms (cf. Table 7.4): the number of correctly predicted OHS_600 spellings (511) achieved by the best combination of rules (including the two analogical variables discussed earlier) is more than twice as large as that by the weakest analogical variable (247). Conversely, the best analogical variable (right constituent family

Table 7.4 *Number of OHS_600 compound spellings predicted correctly by rule-based algorithms and analogical variables*

		Correct O	Correct H	Correct S	Correct Total	Correct Total %
Rules	Comprehensive algorithm	157	187	167	511	85.2
	Maximally efficient algorithm	155	178	151	484	80.7
Analogy	Right constituent family size	117	82	137	336	56.0
	Right constituent family frequency	85	59	152	296	49.3
	Left constituent family size	101	32	133	266	44.3
	Left constituent family frequency	83	18	146	247	41.2

size) reaches only 69 per cent of the performance of the weaker rule-based algorithm.

While there are good reasons for assuming that analogy plays some role as a cognitive process in language use, all of the foregoing seems to suggest that sometimes rules perform better than the kind of analogy which is based on lexemes that are similar due to shared sequences of sounds or letters.

A very important criticism advanced against rules in the past has been that strict rule-based systems do not permit variation (cf. Krott et al. 2001: 80–83). However, rules and variation are not necessarily mutually exclusive, as is demonstrated by the approach described here. While the spelling algorithms in Chapter 6 consist in the sequential consideration of specific values for particular features (which comes very close to a rule-based system positing that a particular type of output is to be assigned e.g. to all adjectives), one should not overlook the fact that the algorithm is actually probabilistic in nature: for a particular combination of features, each spelling variant is assigned a probability based on the target compound's values for those features, and the system always selects the spelling variant with the highest probability. If the values are very similar, this is indicative of a certain degree of uncertainty (whose consequence might be the acceptability of the spelling variant ranked second for a considerable proportion of language users).

More importantly even, the difference between rules and analogy is not as clear-cut as has frequently been suggested, and analogical principles can be reconciled with the feature-based approach of the present study's algorithms by taking into account that the features used in the algorithms were derived from a large set of exemplars. As a consequence, "rules are essentially analogical in nature" (Krott et al. 2001: 53), since e.g. "all compounds with the same second constituent will also have the same number of syllables, frequency, synset count and positional family size" (Bell and Plag 2012: 517).

With regard to the spelling of English compounds, it is noteworthy that those variables which emerge as statistically most significant in determining spelling variant selection (namely part of speech of the compound and length of the compound measured in syllables) concern the compound as a whole rather than its constituents. This explains the low prediction accuracy for the analogical variables: since the analogical constituent family data are always based on a single constituent, they only provide an analogical measure for one part of the compound rather than for the full lexeme. This may lead us to conclude either that the spelling of English compounds is better modelled by rules than by analogy – or, alternatively, that it is based on a relatively abstract type of analogy: while the length criterion may still seem acceptable as an analogical criterion in the sense of a similarity in the holistic formal *gestalt* (cf. e.g. Smith 1988) of the target compound and other compounds in the English language, the part-of-speech criterion stretches the category of analogy considerably. However, it is possible to combine both approaches by recognising analogical phenomena within certain subgroups of compounds only: for instance, analogy in English compound spelling is not necessarily related to all formally similar compounds, but only to the subgroup of family members belonging in the same part-of-speech category as the target compound. This restriction of analogy is driven by the fact that most accounts so far have merely considered nominal noun+noun compounds, so that the compounds investigated in the literature had a more similar structure than the compounds in the present study.

7.4 Cognitive Perspectives on English Compound Spelling

Besides prototype theory and analogical accounts, there are a number of other cognitive perspectives from which the phenomenon of English compound spelling can be considered, e.g. why English has so many spelling variants, whether the pedagogical algorithm has cognitive reality and what an ideal compound spelling system might look like.

One question that emerges when considering rules vs. analogy as cognitive principles that potentially underlie English compound spelling is whether these principles are derived in an analogical way from the material every time that a compound is spelled or whether they are somehow stored and applied in a more abstract way. In view of the immense storage capacity of the human mind (cf. Jurafsky 1996: 114), yet another strategy needs to be considered in addition, namely the storage of individual compounds' spelling. While Bybee (1998: 433) posits that "[a]ctual language seems to rely on memory much more than on abstract analysis", this does not preclude the simultaneous storage of generalisations – because "[i]f linguistic memory is like memory for experience in other domains, it is unlikely that specific instances are completely discarded once a generalization is made" (Bybee 2010: 15). Most likely, therefore, the spelling of English compounds is based on a number of cognitive processing mechanisms which spellers may rely on for different types of compound:

- **Word-specific memory** (which is based on previously encountered stored exemplars) at least
 - a) for the compounds in each spelling category which generalisations are based on (because if these compounds were not stored in memory at least initially, the other compounds could not be related to them)
 - b) for unusual compounds (e.g. those contradicting a language user's intuition or compounds with unusual spellings, such as those with a slash)
- **Subconscious analogy** to similar compounds (in a very wide sense, including principles which emerge from similarities between stored compounds)
- Reliance on **consciously known rules** (e.g. remembering that hyphenations are commonly evaluated in a negative way) when uncertainty reaches the level of awareness.

This model shows a continuum between grammar (in the sense of systematic principles) and the lexicon (in the sense of stored entities), which we can observe in many different areas of language, such as valency grammar (e.g. Herbst et al. 2004; Herbst and Schüller 2008), construction grammar (e.g. Goldberg 1995, 2006) and cognitive grammar (Langacker 2008). Which route is used will depend on a variety of factors:

- For high-frequency compounds which are usually spelled in one particular way only, the activation patterns should be relatively unequivocal and lead to the fast and often subconscious selection of the variant with the highest level of activation. If such compounds are encountered in an unusual spelling, by contrast, this may reach the level of conscious awareness due to the salience of the spelling variant.
- High-frequency compounds exhibiting a large degree of variation in the language will achieve similar frequencies for two or three of the variants in the mental corpus storing previously encountered linguistic use (cf. also Taylor 2005: 3), which can be expected to result in conscious uncertainty corresponding to a position close to the axes of doubt in the multidimensional prototype model in Section 7.2. Depending on the importance attributed to appropriate spelling in the context of writing, the language user can then select from a number of strategies in order to overcome this uncertainty (cf. the beginning of Chapter 6).
- The situation for low-frequency compounds is presumably similar, only that the strength of the prototype will be lower (i.e. the compound will be closer to the converging point of the three variant lines in the multidimensional prototype model): since it is the ratio between the variants rather than their absolute frequency that is important for variant selection (Morton 1969: 167), low-frequency compounds with a clear distribution should be unproblematic, whereas those exhibiting a comparatively large amount of variation in the language and in the mental store should pose the same type of problem as the high-frequency compounds discussed earlier.
- If language users want to spell a compound neologism that they have not read before, they cannot use any stored exemplars in order to retrieve the most likely spelling and need to rely on some other strategy.

The selection process may involve situations in which a speller hesitates between two spellings but has a very strong intuition that the third variant is not appropriate. An interesting question in this context is how such constraining intuitions arise. Two explanations offered by Tomasello (2003: 178–179) and Suttle and Goldberg (submitted) can be applied to compound spelling: if a language user has exclusively encountered a particular spelling variant with one very specific type of compound so far, its usage is *entrenched* and the speller is unlikely to extend that spelling to other types of compound. Alternatively, language users seem to avoid a particular spelling for a specific compound if they have repeatedly witnessed that compound with other spellings. According to Suttle and

Goldberg (submitted), such *statistical preemption* is the stronger explanation of the two. While one cannot exclude particular spellings based on their non-occurrence in a particular sample, the failure to find a particular spelling variant in a very, very large corpus represents a very strong argument for its peripheral status in the language. As a consequence (and in contrast to what is often believed), corpora that are large enough actually provide negative evidence (Stefanowitsch 2006: 62), and it is possible to make a distinction between “accidentally absent” and “significantly absent” structures (Stefanowitsch 2006: 62–63) by using four frequencies: that of a word in a particular slot of a construction, that in all other constructions, the frequency of all other words in the slot of that construction and the frequency of all other words in the corresponding slot of all other constructions (Stefanowitsch 2006: 62–63). Applied to English compound spelling, that would correspond to

- the frequency of a particular compound with a particular spelling (e.g. solid)
- the frequency of the same compound with the other two spellings (open+hyphenated)
- the frequency of all solid compounds minus the frequency of the solid spellings of the target compound
- the frequency of all open and hyphenated compounds minus the frequency of the open and hyphenated spellings of the target compound.

It is thus possible to apply Stefanowitsch’s construction grammar approach to compound spelling in spite of the fact that the categorisation of the three major compound spelling variants as constructions is problematic (since their form side is restricted to the written modality, their meaning/function is difficult to describe, and differences in form do not automatically correlate with differences in meaning). The result does not represent an absolute categorical decision but rather consists in the expression of “a strong statistical tendency” (Stefanowitsch 2006: 70). This is in line with the present, usage-based approach, which relies on probabilistic rather than absolute cues (cf. also Behrens 2007: 208) and recognises that any model of English compound spelling which reflects the linguistic behaviour of the speakers needs to incorporate uncertainty.

If an orthographic system involves variation, as compound spelling does, this raises the question how variation is represented in the mind. By analogy to McQueen and Cutler’s (1998: 409) discussion of morphology, there might be one entry in the mental lexicon which comprises all

spelling variants as possible formal realisations. Those variants which have been encountered before would then have a higher level of resting activation (cf. e.g. Morton 1969: 167) – which could explain why in the reading process, all spelling variants lead to the same entry in the mental lexicon, whereas in compound production, one spelling is easily subconsciously chosen for most compounds. In alternative models of the mental lexicon (McQueen and Cutler 1998: 409), there could be separate entries for each spelling variant linked to each other at a main entry, the nucleus – or one may envisage a network model consisting of a node with the shared information (in this case, the form and meaning of the constituents) linking separate entries. However, there is still a lack of experimental studies which could reveal whether there are several representations of the same compound with different spellings or a single entry with several formal realisations in the mental lexicon.

Yet another question derives from the similarity observed in the empirical study between the predictive accuracy achieved by the two English native speakers and the CompSpell algorithm (cf. 6.2.1): while this may suggest the conclusion that experienced spellers consciously or subconsciously use that algorithm when spelling English compounds, this hypothesis has to be refuted: of the 116 compounds predicted incorrectly by the algorithm, only 30 overlapped with those representing a problem for Participant 1, and only 35 of the algorithm's incorrect predictions were shared by Participant 2. As a consequence, even if the two native speakers were using some kind of strategy, it cannot have been identical to the spelling algorithm. Furthermore, the exclusive subconscious use of an algorithm like a decision tree, which always arrives at precisely one spelling (the most likely option), would contradict the hesitation that human spellers occasionally claim to experience. Nonetheless, the structure of a decision tree comes close to Green and Mehr's (1997: 224) observation that "[m]odels of 'bounded rationality' recognize that human decision-makers have limited ability to attend to multiple cues and that they process information sequentially rather than integratively". They argue that individuals use simple cognitive strategies which "employ sequential evaluation of small numbers of cues", and that such "probabilistic mental models" achieve high accuracy "under conditions of uncertainty or limited information" (Green and Mehr 1997: 224; cf. also Chapter 6). However, one should not treat all language users indiscriminately and overlook the fact that language users differ in their degree of linguistic competence (which may reflect differences in language processing, such as storage or retrieval difficulties; the tendency to devote less attention to form; or smaller and

thus less representative numbers of stored exemplars due to less extensive reading). One may assume that advanced spellers experiencing uncertainty make use of more sophisticated general strategies relying on similarity to the spelling of other compounds (either in the sense of analogy or emerging rules), whereas less advanced spellers might more commonly resort to guessing.

The comparison between the two participants in the native speaker study on English compound spelling (cf. 6.2.1) also yields interesting results: while the spelling of the same eighty-nine OHS_600 compounds was predicted incorrectly by both subjects, this was complemented by an idiosyncratic list of sixty-two incorrectly predicted compounds for Participant 1 and of ninety-four for Participant 2 – which suggests a large degree of individual variation even between experienced spellers. Unexpectedly, the two participants even displayed opposite tendencies in the incorrect spelling variant chosen instead of seventy-three hyphenated OHS_600 compounds (e.g. the noun *by-product* or the adjective *colour-blind*): Participant 1 slightly preferred solid spelling (thirty-nine items = 53 per cent) and Participant 2 overwhelmingly used open spelling (sixty-one items = 84 per cent). These results are in line with recent research by Dąbrowska (2012: 219) which emphasises that “adult monolingual native speakers of the same language do not share the same mental grammar”. Language users may differ with regard to the linguistic cues from the input that are used in the construction of a mental grammar – which results in different grammatical rules – or in terms of the level of abstraction of the emerging rules (which may lead to the same output or not; cf. Dąbrowska 2012). Transferred to the spelling of English compounds, this insight might explain some discrepancies between language users: as pointed out earlier, a large number of variables correlates significantly with spelling variant selection. For instance, head-final morphological structure correlates with open spelling, a verbal first constituent correlates with hyphenation and stress on the first constituent correlates with solid spelling. As a consequence, it is conceivable that different language users attend to different cues (all of which have some statistically significant effect) or weigh the features differently (e.g. by assigning more importance to morphological structure than to the length of the constituents or vice versa), which corresponds to their location at different places in the idiosyncratic decision trees. If, for example, a language user consciously or subconsciously considers word stress more important than part of speech, that person will presumably spell the adjective *candy+striped* solid (because of its stress on the first

constituent) rather than hyphenated (because of its part of speech). In addition, personal (dis)preferences for the spelling variants may act as a conscious filter mechanism. For instance, one of the two native speakers admitted a dispreference for hyphenation, which is comparable to other personality-dependent tendencies and may explain inter-individual differences in spelling variant selection to a certain extent. The present study's spelling algorithms, by contrast, use a statistically determined ordering of the variables which best reflects the data. One may, therefore, conclude that even if cognitive algorithms (rather than e.g. word-specific spelling memory) should underlie the spelling of English compounds at least part of the time, no single algorithm can exhaustively model the cognitive processes involved in English compound spelling due to interpersonal variation in language users' preferences and strategies.

The observable variation renders the system of English compound spelling relatively complex, because only some but not all variants appear acceptable for particular compounds or contexts. An important question that arises is therefore why none of the three main spelling alternatives has been eliminated so far. The diagram in Figure 7.1 can serve to provide an answer to that question: from a logical perspective, the assumed 'ideal' solution combining the features of open, hyphenated and solid compounds would correspond graphically to the more or less triangular field at the centre of Figure 7.1. However, a closer look at the features reveals that these cannot easily be combined: the conveying of unit status (which corresponds to the ellipsoid two-cornered segment on the right) is only compatible with the immediate indication of an entity's constituents (which corresponds to the ellipsoid two-cornered segment on the left) if some discernible item is used that does not break up the chain of letters (cf. Bredel 2011: 34). The immediate solution for a compromise seems to lie in the overlap zone to hyphenation, which is represented by the more or less heart-shaped segment at the bottom of the circle representing this particular spelling variant, and which comprises both analyticalness and unit status. However, the unmarkedness shared by open and solid compounds (which corresponds to the vertical ellipsoid segment at the bottom) is incompatible with hyphenation as long as the latter continues to be widely perceived as a marked and frequently proscribed variant (e.g. in style guides). As a consequence, it is hard to conceive of a spelling which could combine the functions of all three spelling types in its form, and that is presumably also the reason why all three

variants have persisted up to the present.³ The use of a single spelling variant for all compounds would presumably result in decoding disadvantages or decreased efficiency for the spelling of a more or less important subgroup of compounds. From an economical viewpoint, for instance, open and particularly solid spelling are less costly than hyphenation, because they save ink (e.g. if writing on paper) and/or tools (e.g. by wearing out a pen or chisel to a smaller extent) and/or physical effort (particularly obvious if carving texts into stone). By contrast, if one considers which type of spelling differs most from the other two, then at least some cognitive studies (e.g. de Jong et al. 2002: 565) seem to imply that solid and hyphenated compounds, which both convey word unity, are processed in a more similar way than open compounds. As a consequence, one could argue that it would be more economical to dispense with one of those two quasi-synonymous spellings rather than with open spelling. If both economical arguments are combined, hyphens seem most easily dispensable. However, solid spelling has the disadvantage of resulting in very long chains of characters (which might make segmentation more difficult), and open spelling can be problematic as soon as grammatical boundaries are blurred (cf. 5.6), since that makes word unity less easily recognisable. Altogether, it is therefore difficult to argue convincingly in favour of a ban on one particular compound spelling variant – even if it seems that the hyphen's intermediate status makes it a slightly easier target than the other two types.

If one were to attempt the design of an optimal English compound spelling system, a possible alternative to the restriction to a single compound spelling type could be the official acceptance of all three spelling variants for all compounds: if every solution was appropriate, this would also be a very simple principle for the users (even though it would suffer from the disadvantages described earlier for individual spellings). In the current state of the English language, however, there are no generally recognised compound spelling standards, and it is precisely the confusion about the question whether there is only one preferred (and thus more appropriate) variant or whether two or all three are equally likely for a particular compound in a particular context that makes the spelling of English compounds inconsistent and thus difficult (cf. also Hanks 1988).

³ This does, however, not exclude the possibility that other spelling variants could be introduced in spite of the disadvantages which the introduction of additional types might have, or that one or more variants could be dropped at some time in the future in spite of their advantages.

All of this raises the question whether it would be more convenient to have a rigid or a flexible compound spelling system, and speed of acquisition, proportion of ‘correct’ spellings produced by the users and efficiency could be used to evaluate the systems. Kuperman and Bertram (2013: 954) assume that “exposure to a compound in multiple formats will leave weaker traces in the (orthographic) memory for any given format, as compared to the trace left by the same amount of exposure to an orthographically invariable compound”. While an ideal spelling system for English compounds may involve the consistent use of different spelling variants for specific types of compound (like hyphenation for adjectives), variation in the spelling of individual words is a disadvantage: even if the coexistence of several forms for one compound results in the appropriateness of more than one solution, the processing time required by the resolution of uncertainty (particularly when it takes place at the conscious level) incurs a greater cost in a system in which at least some spellings are uncommon than if each compound is consistently spelled in one particular way only. This may also explain Sebba’s (2007: 153–154) observation that the public perceives optionality as very problematic and rather wants to be told how to spell (Sebba 2007: 153–154).

However, that does not mean that the present system should necessarily be changed. If we also consider economy as a variable of influence, any major change in the system with the ensuing need to learn new conventions would result in increased effort – and thus in a disadvantage in economical terms – for the language users already familiar with that system. Furthermore, keeping a non-optimal spelling for individual compounds may be less problematic than one might initially believe, since it seems “that readers are able to optimise their processing strategies . . . over time when a compound consistently appears in the same format” (Kuperman and Bertram 2013: 942).

While spelling is an aspect of language that is partly shaped by “cultural norms” (Wiese 2004: 326; but cf. also Chapter 3 on the role of language users for emergent norms), it is interesting to explore what a cognitively optimal compound spelling system might look like. An important distinction that needs to be made in this context is whether we are dealing with language comprehension or production, since an optimal system for coding is not necessarily also an optimal system for decoding (Wiese 2004: 309). On the coding side of English compound spelling, there do not seem to be many psycholinguistic studies, which leads Kreiner and Gough (1990: 103) to conclude that cognitive spelling research does not receive the treatment in the literature which it deserves. By contrast, several

psycholinguistic studies on compound spelling have been conducted from the reading perspective (cf. 1.1.1). Provided that their findings based on different languages and different types of compound can be generalised, we can draw the following conclusions:

- **open spelling**
 - speeds up lexical decision (Juhasz, Inhoff and Rayner 2005)
 - facilitates access to the constituent lexemes (Juhasz et al. 2005; cf. also Sanchez 2008: 263–269; Sandra 1990; Fiorentino and Fund-Reznicek 2009).
 - renders the processing of compounds more comparable to that of simplex words (de Jong et al. 2002)
- **hyphenated spelling**
 - benefits early processes and disrupts later ones (Bertram et al. 2011)
- **solid spelling**
 - benefits the specification of full compound meaning (Juhasz et al. 2005)
 - leads to fewer refixations (Juhasz et al. 2005).

This suggests that all three spelling variants have their justification due to the differences between them, and that from a cognitive receptive perspective, hyphenation is more similar to open than to solid spelling (in spite of the fact that it also constitutes an uninterrupted sequence of characters).

Furthermore, the effect of boundary markers was tested in psycholinguistic studies. Epelboim et al. (1997) inserted different types of filler into texts and found that this slowed down reading considerably. How spaces were filled resulted in a different strength of effect: shaded boxes (ASCII character 176) slowed down reading least; next came digits (neither <1> nor <0> were used because of their resemblance to the letters <I> and <O>), then Greek letters and then Latin letters. Since this suggests that the effect of the fillers increases with their resemblance to the standard letters used in a language, one may conclude that the introduction of hyphens into compounds should be particularly unproblematic: the short horizontal line is clearly distinct from letters, symbols and even most punctuation marks (which usually have a vertical rather than a horizontal dimension).

Since texts are usually only written once (even if there might be revisions) but may be read far more often – from the note on a fridge door to

a best-selling novel read millions of times – user-friendly decoding should have highest priority in an optimal spelling system. However, it seems that compound comprehension is rarely hindered by compound spelling even in the present system: users seem to focus on the form of compounds only when there is a strong clash with the expectations, e.g. when reading constructed sentences such as *‘It was not very farsighted of our local agony-aunt to celebrate her birth-day with so many cock tails*, in which one would usually expect hyphenation in *far-sighted*, open spelling in *agony aunt* and solid spelling in *birthday* and *cocktails* (as in the OHS_extra list). The unproblematic decoding compared to the occasionally difficult coding situation (in which the speller is required to select some variant in spite of uncertainty) is possibly due to the fact that the formal difference between the three spelling variants is relatively inconspicuous, and since only the hyphen has a material physical representation, the distribution may go unnoticed for less attentive readers.

Note that all of the foregoing discussions are based on the implicit assumption that the selection of a particular compound spelling variant is intentional and meaningful, at least in some cases. However, one may call into question to what extent Murray’s (1997: 126) concept of *agency* (which she defines as “the satisfying power to take meaningful action and see the results of our decisions and choices”) can actually be applied to the spelling of English compounds (cf. also 3.1). For instance, there has been a long tradition of re-editing texts belonging in the canon of classical literature by changing their punctuation for better readability in print (Nebrig and Spoerhase 2012: 13–16). This would seem to imply that punctuation is not regarded as part of the poetic message of the author, and in the majority of printed books, it is therefore difficult to determine who is responsible for the punctuation – the author, the copy-editor, the printer or all three (Parkes 1992: 5). However, while some authors do not attach too much importance to their punctuation and allow or even ask their publishers to correct their punctuation (e.g. William Wordsworth and Charlotte Brontë), others use it very intentionally (e.g. Charles Dickens, Heinrich von Kleist or Gertrude Stein; cf. Nebrig and Spoerhase 2012: 30; Parkes 1992: 5). It is already difficult to evaluate the extent to which professional writers deliberately select particular spellings and how much meaning they attach to these, but the task becomes almost impossible for compound spelling behaviour in general. Since the intentionality of most individual compound spellings should not be overestimated, the present approach therefore analyses averaged preferences.

7.5 Integrating Change

According to Chambers and Trudgill (1980: 145), “throughout the history of linguistics, linguists have tended to act as if language were not variable,” and “[m]ost linguistic theories have started from the assumption that variability in language is unmanageable, or uninteresting, or both”. One reason for that may be that “since Saussure, variability has been defined as lying outside the linguistic system, external to *langue*, competence, and grammar” (Guy 2011: 179). At the same time, however, it is commonly acknowledged that change is a substantial quality of natural languages (cf. e.g. Keller 1994: 21; Rauch 1989: 376). Variation and change cannot be separated from each other, because “variation is the synchronic face of change, and change is nothing more than diachronic variation” (Rauch 1989: 376). Even if the coexistence of variants over several centuries “suggests that change is not an inevitable outcome of variation” (Guy 2011: 179), we can conclude that any theory of linguistic variation needs to be compatible with language change. Furthermore, norms for living languages cannot be fixed once for all times; they are in perpetual flux (cf. Hundt 2010: 45), and it is therefore necessary “to reconcile the static quality of a fixed standard with the dynamics of social and linguistic change” (Johnson 2002: 566). In contrast to Saussure’s (1916/1959: 27) assumption that “language is constantly evolving, whereas writing tends to remain stable”, it may be that aspects such as compound spelling or punctuation, which often pass unnoticed but occur very frequently when considered on a general level, are actually particularly likely to develop in a way that is favourable from a cognitive point of view. This would correspond to Mair’s (2009: 3) observation that comprehensive and far-reaching changes – e.g. the use of *-ing* forms in previously infinitival functions – “tend to go unnoticed” if they “proceed below the level of conscious speaker awareness and hence do not arouse prescriptive concerns” – in contrast to the allegedly ongoing disappearance of *whom*, which is discussed extensively in public.

Since language is permanently subject to change, the intuition of its users needs to adapt to changing usage from time to time. Interestingly, English compound spelling is an area in which change is actually expected: thus Reiser (2007) advises her readers to make sure that the dictionary they consult in case of doubt is current, and Wilbers (1997) even suggests to use a dictionary that is replaced “every five years to keep abreast of vanishing hyphens”. It seems that change in this area is so certain that language users risk not being up to date if they do not keep an

eye on ongoing developments. This is a very different situation from other aspects of language, where hard fights are fought in order to maintain the status quo, and where change is usually equated with decay (cf. Aitchison 1991: 7).

Change as a universal feature of language can be accommodated by the present account of English compound spelling via the consideration of changes in the variables that play a role in the spelling of individual English compounds. Some of these may change relatively easily (e.g. a compound's degree of idiomaticity due to semantic change, or the increased frequency of a compound or constituent due to the intense public discussion of a relevant event), with the corresponding effect: if a particular noun compound with variable stress tends to be stressed increasingly on the first constituent, this may correlate with an increased tendency towards solid spelling, as observed in the present study (cf. Table 5.31). Other variables, by contrast, cannot change that easily: either a morphological structure is head-final or not, and reinterpretations are presumably very rare. Similarly, compound-final *-ing*, *-ed* and *-er* are either present or not; a change in this variable would require some kind of merger of base and suffix, but this rarely happens. If we merely consider the most influential features, which are retained in the CompSpell algorithm (i.e. part of speech, word length and constituent length), these offer hardly any opportunities for change. However, if many compounds in the constituent family of a particular compound change their spelling due to these reasons (or others), the target compound may exhibit an increased tendency towards that spelling as well. This would be observable as ever more usage tokens with the new spelling, in spite of no changes in the qualities of the compound itself – even though the overlap in one constituent makes it very likely that changes may be related to some quality of that constituent and thus apply to the target compound anyway. The apparent contradiction between the seemingly immutable algorithm on the one hand and the requirement for change on the other hand can also be resolved by considering that the algorithm is only an efficient shortcut to the spelling preferences which emerge based on a variety of factors. The previous sections showed that a wide range of variables correlates with the preferences or dispreferences for particular spellings. Changes are thus possible, but only indirectly, by considering what characterises the central exemplars within each of the three spelling categories at some time in the future. If the prototypical compounds with exclusively open, hyphenated and solid spelling have changed by then, the most efficient algorithm may look somewhat different, but the stock of features from which it is composed may still be the same. This could be

determined by a follow-on study after an appropriate amount of time. However, prototypical spellings presumably change slowly, due to their high level of resting activation compared to alternative spellings: Bybee (2003: 42) suggests an exemplar-based model of the mental lexicon, in which the individual variation encountered in usage tokens is represented at each entry, as in a corpus. The most frequent variants constitute the centre of the variational range and the prototype for the category. This centre “may gradually shift as words are used” (Bybee 2003: 42), which has an important consequence for language change: it is very likely that – in the beginning of a spelling change – language users still encounter the conservative spelling variant to a considerable extent, e.g. because of its presence in books at school, in public libraries and on the bookshelves at home. This effect is increased if one or more influential newspapers or other written media do not adopt the new spelling variant. All these tokens of usage strengthen the conservative spelling variants in the mental lexicon and slow down spelling changes. If the likelihood for old and new spelling is very similar, the most recently encountered variant may exert a priming effect and induce the language user to select the next compound spelling accordingly. This may explain the strong hesitation of language users in periods of transition. The existence of variant spellings may thus be attributed

- a) to similar frequencies of occurrence in the language and consequently also in the mental lexicon
- b) to contradictory principles derived from current usage
- c) to language change in progress.

Since the spelling algorithm predicts 75 per cent of compounds correctly, the remaining 25 per cent might leave room for a certain shift in spelling preferences without requiring a complete remodelling of the algorithm in the very near future.

This raises the question what factors influence change in language (cf. also Sanchez-Stockhammer 2015) and, more specifically, in compound spelling. While the question cannot be answered exhaustively, since new variables may always emerge, the successful introduction of new media seems to play an important role and may result in new spelling preferences, because a particular spelling variant is favoured as more efficient (cf. later in this chapter). Furthermore, the repeated perception of compounds from another language with varying conventions can favour the adoption of a new spelling format. Widely used spellcheckers which do not properly accommodate British English spelling may result in a tendency to adopt American English spelling preferences (although the differences regarding

compound spelling are only minimal). Furthermore, the strong influence of Anglo-American culture in Germany and the resulting high frequency with which English compounds are being read (e.g. in advertisements) seems to have fostered a trend towards hyphenated and (traditionally incorrect) open spellings in German compounds. On a general level, Schneider (1997: 50) posits that “[n]atural, internal forces trigger change; social, external factors determine its progress and embedding”.

While the Neogrammarians in the nineteenth century expected sound changes to affect all words of a language simultaneously, the current view is that “in the majority of cases, a change affects different words at different times” (Aitchison 1991: 76–77; cf. also Lass 1997: 140). That also applies to compound spelling. If linguistic change is represented in a diagram with time on the *x*-axis and change on the *y*-axis, it assumes the shape of an S-shaped curve (cf. Aitchison 1991: 83–86; Graddol 1997: 18; Lass 1997: 370 for the following): at first, an innovation tends to affect only a few words or constructions. This is followed by a phase in which many items are affected within a very short period of time – and it is usually during this stage of rapid change that language users become aware of the ongoing developments. For example, the increased frequency of a particular spelling variant in positively evaluated texts may result in a snowball effect and reduce the reluctance to use that spelling in other texts (cf. Vallins 1954: 192). At that stage, observers tend to expect that the trend will go on indefinitely, but actually, the process then slows down. The last few items either change only slowly (although one cannot predict with what speed; cf. Bauer 1994: 25), until the change achieves a “natural end point” (Graddol 1997: 18), or not at all, since “[c]hanges may never complete, but may abort at virtually any stage” (Lass 1997: 140). Bauer (1994: 25) concludes that “observation of a change in progress is not a sufficient basis for making a prediction about the outcome of that change” and even mentions the possibility of a reversal of the change (cf. also Nevalainen 2015). Making predictions about linguistic change is therefore problematic, particularly since extralinguistic developments (such as technical advances), which may drastically change ongoing developments, cannot be foreseen. Nevertheless, if all of this is considered jointly with the results of the empirical study, a few tentative predictions can be made regarding future developments in the spelling of English compounds. However, these predictions only express tendencies and are furthermore linked to the prerequisite that no completely new variables enter the system:

- **It is very likely that the spelling of English compounds will become more regular.**

A few centuries ago, few people in Europe were able to read, and even fewer actively wrote (cf. also Milroy and Milroy 1985: 50). Most of the latter were specialised professionals, such as the scribes in the Middle Ages, or – a few centuries later – authors or journalists who knew that they could rely on a copy-editor for their publications. It is only in the past few decades that an increasing number of ‘normal’ language users have been writing an increasing number of texts in English, thereby also increasing the number of compound tokens to be spelled. This development has gained particular momentum ever since the spread of personal computers, emails and particularly the web 2.0 and its successive stages with manifold opportunities for written interaction. In view of this increase in frequency, one may expect the increased regularity which is commonly observed in such contexts: thus second-generation sign language users tend to regularise their parents’ system (cf. Real and Griffiths 2009: 317), and experiments with human subjects and computer programs by Real and Griffiths (2009) and Kirby, Cornish and Smith (2008) find that linguistic structures in general tend to regularise over time. They observe that the outcome of iterated learning (“a process in which an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way”) undergoes an evolutionary “invisible hand” process “leading to phenomena that are the result of human action but are not intentional artifacts” (Kirby et al. 2008: 10681).

- **New compounds will presumably not develop from open via hyphenated to solid spelling as a rule.**

It was shown in the empirical study that such a process could not be observed in the past, and since there is no support for it in the present, there is no reason to assume that it will play a role in the near future, either. As a consequence, we should not expect all compounds to reach the stage of solid spelling sooner or later. Since verbs were never spelled open in the empirical study, not even in the OHS_extra data, this is very unlikely to change in the near future. Similarly, the almost complete absence of open spellings among the adjectives and adverbs speaks in favour of the preservation of these patterns as well. As a consequence, Hale and Scanlon’s (1999: 13) recommendation of using solid spelling when in doubt in order to “[s]ave a keystroke” and to anticipate future spelling developments in digital writing must therefore be refuted, since

it does not take part of speech into account and also contradicts the next expectation:

- **An increasing number of noun compounds will presumably be spelled open.**

‘Normal’ language users are typing an increasing amount of text now compared to a few years ago (e.g. emails, blog postings, text messages, social network posts). Probably, the future will see even other types of currently unthinkable written communication. If the need to type ever more text is combined with the observation that increased speed of typing results in a tendency towards open spelling, we may assume that noun compounds as the only part-of-speech-based category which permits open spelling (cf. 6.3) will tend slightly towards open spelling in the future. The view that different linguistic principles can be most efficient at different times due to changing external conditions goes back to the functional view of language change, which posits that “[l]anguage alters as the needs of its users alter” (Aitchison 1991: 117): the historical transition from loud to silent reading resulted in the increased use of punctuation marks (Parkes 1992: 1), and in the digital age, the need to save space has been replaced by the need to permit quick processing. However, if typing compounds should be superseded by voice recognition or some other method as the most important way of realising compounds in the written modality, this may change the picture completely. Most likely, this would result in a standardising and conservative tendency, since such software would presumably rely on dictionary-derived word lists (possibly combined with a few syntactic restrictions) and reduce the amount of variation caused by individual spellers.

7.6 Summary

This chapter considers how existing theoretical and cognitive approaches can be used to model the spelling of English compounds. In contrast to previous research on English compound spelling (which focuses on variation), the present approach considers those English compounds which do not vary in their spelling as central exemplars of the three spelling variant categories OPEN SPELLING, HYPHENATED SPELLING and SOLID SPELLING, respectively. Within this prototype-based model, compounds exhibiting variation can be considered members of more than one category, and depending on their frequency and proportion of occurrence in each of

the three formats, specific English compounds are situated more or less closely to the centres of the categories. This chapter discusses different ways of representing this graphically by using concentric circles and overcomes the shortcomings of two-dimensional models by reinterpreting the arrows signalling increasing strength of category membership as the axes of a coordinate system. The resulting three-dimensional prototype model of English compound spelling can easily be expanded into a multidimensional model capable of dealing even with the enormous amount of variation that is potentially present in compounds with more than two constituents. In a more traditional application of prototype theory to English compound spelling, we can observe that the most prototypical English compounds (i.e. biconstituent, lexicalised, head-final nominal compounds consisting of noun+noun with stress on the first constituent) are mainly spelled solid, with open spelling coming very close, whereas the largest number of compound types and tokens in the data of the empirical study is spelled open. This seems to suggest a twofold prototype – which may explain the difficulties that many language users experience with English compound spelling.

This chapter also discusses the explanatory power of rules and analogy for English compound spelling. When comparing the performance of the empirical study's analogical variables with the predictions made by the maximally efficient spelling algorithm doing without them, the algorithm outperforms analogy with 80–85 per cent as against 41–56 per cent correct predictions. This is presumably due to the fact that traditional analogical accounts are merely based on features which are related to individual constituents but not to the whole compound (such as part of speech, which is a particularly strong predictor for English compound spelling). Nonetheless, the opposition frequently set up between rules and analogy can be reconciled by taking into account that the features used in the rules have been derived from a large set of exemplars or by considering analogical principles that only apply to specific subcategories (e.g. noun+noun compounds).

The discussion of English compound spelling from a cognitive perspective centres on various questions, such as the processing mechanisms used for different types of compound (word-specific memory, analogy and rules; depending on the compound's high or low frequency and its small or large variation in spelling usage), the representation of variation in the mental lexicon and the cognitive plausibility of the pedagogical algorithm, particularly in view of the individual differences between language users and potential variation in the intentionality of spelling variant selection.

Another issue discussed in some detail is the question what a compound spelling system that is optimal from a cognitive point of view might look like. This chapter attempts to find an explanation for the observable persistence of variation and considers the advantages and disadvantages of rigid vs. flexible spelling systems in the light of the discrepancy between the comprehension perspective (which may easily accept several spellings in many cases), and the production perspective (in which a single spelling variant needs to be selected).

The present study analyses statistical tendencies in the spelling of biconstituent British English compounds at the beginning of the twenty-first century. In view of the universality of language change, this chapter finally discusses how the currently observable principles can be reconciled with the dynamics of linguistic change. The principles derived from the empirical study even permit the tentative formulation of a number of predictions regarding likely developments in English compound spelling (provided that no disruptive forces apply).

Summary and Conclusion

The spelling of English compounds is an aspect of language production which is notoriously difficult for language users at all levels of proficiency and less standardised than other orthographic aspects of the language. This can presumably be attributed to its consideration as less important than other orthographic issues, possibly because no letters are concerned and because misunderstandings due to unusual or even inappropriate compound spelling are very infrequent. Still, anyone writing a text in English is constantly faced with the usually subconscious or (in the case of doubt) conscious question whether to use open spelling (*drinking fountain*), hyphenation (*far-off*) or solid spelling (*airport*) for individual compounds. A few options exist or are conceivable (e.g. the use of slashes in the compound *aural/oral approach*; cf. 2.5.4), but these are so rare that they can be disregarded in the following. Since one variant has to be selected necessarily in the written coding process, the problem can hardly be avoided (only e.g. by using phrasal paraphrases with obligatory open spelling rather than compounds). Since several spelling variants are commonly used for some compounds, whereas strong preferences can be observed for other compounds, a distinction can thus be made between

- a) appropriate spelling (i.e. the selection of the single most frequent variant or of any one among roughly equally frequent alternatives), e.g. *gender bender* or *gender-bender* (with respective frequencies of three and two in the dictionaries used here)
- b) unusual spelling (i.e. the selection of a not completely uncommon minority variant), e.g. *eyeshadow* instead of *eye shadow* (with one as against five occurrences in the dictionaries)
- c) inappropriate spelling (i.e. the selection of an extremely uncommon minority variant in the face of one or more very frequent preferred

variants), e.g. ²*calendarmonth* instead of *calendar month* (which is spelled open in all the dictionaries used here).

Language users may seek guidance in different types of reference work, which all have their advantages and disadvantages (cf. Chapter 1.1 and 6.3): while style guides provide general rules that are applicable to many cases, the retrieval of the relevant information is relatively complex. The same is true of grammars, only that their perspective is descriptive rather than prescriptive. Dictionaries, by contrast, permit easy access to any compound to be spelled, but the advice that they offer is restricted to the list of words that they contain and cannot be transferred to new compounds. All this seems to suggest that a simple compound spelling algorithm with a few exception principles would be a desirable achievement, particularly for the teaching of English as a foreign language. The empirical research forming the basis for the creation of such an algorithm tested three dominating views about the spelling of English compounds in the literature:

- a) The spelling of English compounds is chaotic (e.g. Fowler 1926: 243; Merriam-Webster 2001: 99).
- b) The spelling of individual English compounds develops from an open via a hyphenated stage towards solid spelling (e.g. Quirk et al. 1985: 1537; Partridge 1953: 138).
- c) British English uses more hyphenation than American English (e.g. Quirk et al. 1985: 1569; Butcher 1992: 154).

Possible regional variation was accounted for by concentrating on compound spelling in British English and by comparing it with American English. Note that the first two views expressed earlier are not necessarily mutually exclusive, since a different speed of development of individual compounds may result in what looks like superficial chaos. In order to test these hypotheses and determine whether there are any principles underlying English compound spelling, it was first necessary to delimit the compound concept. The view adopted here is that a compound concept which takes the spelling needs of language users into account needs to embrace a very wide compound concept. As a consequence, the compound concept of the present study goes beyond the nominal noun+noun compounds to which previous research focusing on English compound spelling usually limits itself (e.g. Sepp 2006; Rakić 2009; Kuperman and Bertram 2013). This requires the distinction of compounds from a large number of adjacent and sometimes overlapping categories, such as simplex lexemes, derivatives, other word formations, multi-word items, names and

particularly phrases (cf. Chapter 2). After the evaluation of numerous possible distinctive criteria from the literature, compounds are generally defined as complex lexemes which

- refer to a unified semantic concept
- consist of at least two constituents that occur as free, synchronically recognisable and semantically relevant lexemes each
- contain no affixation on the highest structural level
- can be assigned a joint part of speech
- cannot be interrupted by the insertion of lexical material and
- only once permit the application of each type of inflection to their base form.

Another aspect neglected in previous research on the spelling of English compounds is the normative background of the phenomenon: language users seem to expect that some spellings are right and others wrong. Chapter 3 explores numerous reasons for language users' apparent strong wish to comply with a spelling standard (such as issues of communicative success, power, stability, status and identity) and discusses the role of governments, the publishing business, linguistic experts (including lexicographers) and language users in shaping what is perceived as correct spelling. The conclusion to be drawn is that norms and usage mutually influence each other: modern reference works consider usage in the form of corpora, and the texts contained therein (particularly the edited ones) rely on reference works in cases of formal doubt.

Previous research on English compound spelling mainly focuses on corpus-derived variation and attempts to explain why variation is possible (e.g. Sepp 2006: 10). The approach followed here, by contrast, maintains that absolutely free variation is relatively rare and that one variant is usually more likely than the other two in the spelling of individual compounds. The approach is based on the assumption that there is a common core of English compounds whose spelling is relatively unquestioned and that the principles derived from these prototypical exemplars within the categories of open, hyphenated and solid spelling may in turn be applied to the spelling of all other compounds, so as to determine which spelling variant these are most likely to favour. The large-scale empirical study carried out with the aim of testing this hypothesis used a list of compounds from the *Longman Dictionary of Contemporary English* (LDOCE 2009) which had been coded and extracted by the lexicographers at Longman. Only those compounds which agreed with the present study's compound definition were retained in the Master Compound List comprising more than 10,000

compounds. The invariant compounds for each spelling format were determined by the automatic comparison of the Master List with the English lemma lists (including part of speech) from LDOCE and five other dictionaries: *Cambridge Advanced Learner's Dictionary* (2008), *Macmillan English Dictionary for Advanced Learners* (2007), Langenscheidt *Taschenwörterbuch Englisch-Deutsch* (2007), *Oxford Advanced Learner's Dictionary* (2005) and *Collins English Dictionary* (2004). In addition, corpus data from analogously structured representative corpora for different periods and varieties of English were used: BLOB-1931, LOB (1961), FLOB (1991) and BEO6 (2006) for British English, and Brown (1961) and FROWN (1991) for American English. The diachronic data were complemented by information on the first attested spelling in the *Oxford English Dictionary* (2009). In addition, three specialised corpora were analysed: the Blog Authorship Corpus for blog texts, the NPS Chat Corpus containing material from chats and the CorTxt text message corpus (cf. Chapter 4 for more details regarding material and method).

Chapter 5 investigates a large number of features which may possibly influence the spelling of English compounds. These come from very different domains: some are related to spelling, others to length, frequency, phonology, morphology, grammar, semantics, diachronic aspects, discourse, the system of the English language, and yet others are extralinguistic. In view of the large amount of features which were expected to exert some influence on the spelling of English compounds, the features which could not be coded automatically by the program CompSpell were coded manually for a random selection of 600 compounds with a frequency of occurrence of five or more in the dictionaries (with the corresponding part of speech). The so-called OHS_600 dataset comprises 200 biconstituent compounds each with uniquely open, hyphenated and solid spelling and thus represents the core compounds referred to earlier. This sample is complemented by several other datasets based on the original Master List: OHS_extra comprises the 3,864 biconstituent compounds corresponding to the same criteria as OHS_600 but which were not chosen in the randomised selection process. The Master_5+ list contains all 1,196 compounds occurring at least five times in the dictionaries which are not part of OHS_600 or OHS_extra items, either because they comprise more than two constituents or because they vary in their spelling. Finally, the Master_1-4 list contains the 4,381 compounds from the Master List which occur in only up to four dictionaries with the corresponding part of speech.

The OHS_600 compounds, supported by the other lists, were used to test more than sixty hypotheses on the spelling of English compounds. Chapter 5 presents and discusses the detailed results, comparing them with the findings of previous research. Summaries at the end of each group of features (e.g. phonological variables) provide a quick overview of the most important findings. With regard to the three common views about English compound spelling cited previously, we can draw the following conclusions:

- a) The spelling of English compounds is not chaotic and is actually governed by a large number of variables which are statistically significant when tested individually (cf. Table A.9 in the Appendix). Possibly, this abundance of underlying principles is perceived as a superficial lack of patterns by many language users.
- b) Similarly, no support was found for the hypothesis that compounds start their life open, then go through a hyphenated stage and finally become solid.
- c) Nor could the third general view in the literature be confirmed: in the dataset under consideration, hyphenation was not more common in British English than in American English.

In the next step (cf. Chapter 6), it was attempted to determine the extent to which the combination of all statistically significant variables can account for the spelling of individual compounds. The comprehensive Algorithm 1 (whose initial thirty-four features the program R automatically reduced to eleven due to correlations) predicts 85.2 per cent of the OHS_600 compounds correctly. The user-friendly Algorithm 2 (which uses only the significant and objective variables that are easy to apply for language users without prior linguistic training) performs almost equally well on the training set with a predictive accuracy of 81.3 per cent. If the number of variables is maximally reduced, the proportion of compounds spelled correctly by the single-variable take-the-best Algorithm 3 is very low at 59.2 per cent. By contrast, the predictive accuracy of the slightly more complex, maximally efficient Algorithm 4 (which uses merely the three features part of speech of the compound, length of the compound measured in syllables and length of the second constituent measured in letters) is higher again with predictive accuracies of:

80.7% for	OHS_600
76.3% for	OHS_extra
61.0% for	Master_5+_tendency
72.9% for	CompText (a corpus of British English compiled specifically for the present study; the compounds in the test sample do not occur in the Master List).

Table 8.1 *The CompSpell algorithm*

<ul style="list-style-type: none">• Adjective (<i>broken-down</i>)Adverb (<i>well-nigh</i>)Verb (<i>chain-smoke</i>)	Hyphenated
<ul style="list-style-type: none">• Noun<ul style="list-style-type: none">• three or more syllables (<i>bathing suit</i>)• two syllables<ul style="list-style-type: none">◦ second constituent: up to two letters (<i>close-up</i>)◦ second constituent: more than two letters (<i>coastline</i>)	<div>Open</div> <div>Hyphenated</div> <div>Solid</div>

These results are so similar to the predictive accuracies achieved by the intuitive spellings of two educated native speakers of British English that Algorithm 4 (the CompSpell algorithm) can be considered a very good approximation to native speaker competence for the spelling of biconstituent English compounds.

This algorithm can be expected to predict the spelling of roughly three out of four compounds for all parts of speech correctly. Its simplicity and efficiency make it a desirable complementation of existing material in the teaching of English as a foreign language. More advanced learners (but also native speakers of English) may use the algorithm as a decision instrument in cases where their intuition can provide no guidance. Where the result of the algorithm contradicts more advanced language users' intuition, the consideration of a number of exception and blocking principles (suggesting e.g. open spelling for phrase-like nouns, hyphenation for compounds with heterogeneous constituents or solid spelling for compounds with fore-stress) can provide support for the next most probable spelling (cf. Table 6.16 and Table 6.17).

After these chapters with an empirical and applied linguistic focus, Chapter 7 returns to the theoretical perspective by exploring the relationship between the three major spelling variants and by discussing various ways how the spelling of English compounds could be modelled based on existing theories of language.

In prototype theory, the extent to which a compound's spelling varies determines its closeness to the centres of the three spelling variant categories (open, hyphenated and solid spelling). The limitations of different two-dimensional graphical representations pave the way for a three-dimensional prototype model of English compound spelling. If the frequencies of occurrence for any number of variants are known, these can be

used as coordinates in a multidimensional model capable of dealing even with the potentially very large variation in compounds with more than two constituents. While the most prototypical English compounds (i.e. biconstituent, lexicalised, head-final nominal noun+noun compounds with stress on the first constituent) in a more traditional account are mainly spelled solid (with open spelling coming very close), the clear majority of all compounds in the empirical study use open spelling. This suggests a twofold prototype, which may explain compound spelling difficulties to a certain extent.

Chapter 7 also contrasts word-specific memory, analogy and rules and comes to the conclusion that these are not mutually exclusive, since both analogies and descriptive principles (i.e. rules) are derived from individual exemplars – which need to be stored in the mental lexicon. Furthermore, the chapter explores the requirements of a cognitively optimal compound spelling system and discusses how orthographic variation may be represented in the mental lexicon. A cognitive model of English compound spelling needs to take into account that different (although not all) variants may be accepted in comprehension, whereas in production, one variant is frequently selected very quickly and without conscious awareness because of its preference over the two alternatives. This can be explained by different levels of resting activation for the spelling variants, which need to be linked in some way or another. Depending on the frequency (including zero for new formations) and degree of spelling variation of a compound, the spelling mechanism may involve word-specific memory, analogy or rules. Since all three spelling variants have their advantages and disadvantages, an optimal system may use more than one of these for different types of compound – but it would presumably not vary in the spelling of individual compounds. The final part of Chapter 7 explores how the present study's algorithms, which model the prototypical spelling of biconstituent compounds in British English reference works, can integrate the universal expectation of language change.

The account of English compound spelling presented here is quite unique in that it provides a tool which permits predictions for the preferred usage of present-day English compounds, and whose results can be subjected to empirical testing instead of being limited to a post hoc explanation (like most other accounts). Provided that no disruptive forces apply, it is even possible to make a number of likely relatively general predictions for future developments in English compound spelling, such as a tendency

towards regularisation and an increase in nouns with open spelling in digital texts.

Of course, the study presented here also has its limitations: first of all, it concentrates on typical compounds in one variety of English, namely British English – but for certain aspects, comparisons with American English are drawn. Since the study is based on dictionary data, it can also be criticised for modelling an ideal use of language – but in a domain which is so frequently approached from a prescriptive perspective, that is not necessarily a disadvantage, since it means that the algorithm represents accepted usage. In addition, most dictionaries state explicitly that they use corpora containing language usage as the basis for their lexicographical description of English, and the present approach is furthermore complemented by several corpus analyses. Also, while the algorithm provides a relatively efficient strategy for the spelling of established biconstituent compounds, its predictive accuracy for less common compounds is presumably lower – but then, these compounds are more likely to have several acceptable spellings anyway, which reduces the probability of choosing a completely unacceptable variant, and the CompSpell algorithm achieves relatively high prediction accuracy even for compounds which are not listed in dictionaries (72.9 per cent for the CompText compounds). Another limitation concerning the diachronic part of the study is that the corpus-based investigation of British English is restricted to the period between 1931 and 2006. In order to overcome this limitation, the first spelling in the *Oxford English Dictionary*'s earliest attestations was also coded, but this is not entirely unproblematic in view of the frequent modification of compound spellings in editions of early English texts, which formed the basis for the dictionary quotations. This could, however, be remedied by a corpus study of historical English corpora considering the editing of its material.

To conclude, the present study attempted to fill a gap in the literature by seeking to transcend the individuality of lexical items and to determine the regularities which make the preferred spelling for one compound similar to that for another. Its aim was to “uncover those regularities that appear to be exploitable by a competent speller” (Carney 1994: 4) and to serve “both as a reference guide and for further research into the nature of language itself” (Ghomeshi 2010: 72). Since the CompSpell algorithm is simple and has a reasonable predictive accuracy (even for 77 per cent of the OHS_600 noun compounds), it provides a convenient tool for beginning learners and can support advanced spellers in doubtful cases where a decision has to be made. While its principles appear to be context-free, they are indirectly

context-sensitive via part of speech as a shortcut for certain types of syntactic context. Many of the variables tested in the present study have been mentioned before in reference works of various kinds, but the present study is the first to use a weighted ranking of these specific criteria to arrive at a high-efficiency spelling algorithm.

The spelling of English compounds superficially looks like a very limited and trivial research subject. However, as soon as one starts considering it in more detail, it becomes obvious that it is governed by principles which are also important in other domains of language, that it is actually quite intricate and complex and that it raises a large number of interesting questions. Even if there is still plenty of room for future research (e.g. on the role of establishment in discourse, on compounds with more than two constituents or on the pedagogical value and efficiency of the spelling algorithms), some of these questions have hopefully been answered by the present study: we have covered numerous aspects of English compound spelling, analysing a large number of English compounds following a very wide definition of the concept and using data from dictionaries and from corpora, from different varieties and from different types of text. We have considered historical developments and a multitude of potentially distinctive features to arrive at a maximally efficient combination for present-day British English. We have considered the normative background and discussed how the phenomenon can be modelled on the theoretical level and from a cognitive perspective – all in order to provide a comprehensive treatment of the determinants of English compound spelling.

Appendix

Table A.1 *The 200 OHS_600 compounds with exclusively open spelling*

age group	compact disc	glass ceiling
anabolic steroid	company car	glove compartment
artificial insemination	cooking book	golden parachute
asking price	copy editor	graduate school
baby boom	correspondence course	hard copy
banker's order	cosmetic surgery	hard currency
bathing costume	cough mixture	heavy breather
bathing suit	cover charge	herbaceous border
battle cry	crash helmet	high table
bean curd	crazy paving	high tea
big business	cream tea	hit list
big time	day return	house martin
black pepper	deposit account	housing estate
blood brother	dining car	human resources
blood sport	district court	human right
blue jeans	double chin	incidental music
boarding card	drinking fountain	inorganic chemistry
book token	drum machine	ivory tower
boom town	duffel bag	legal tender
bottle bank	duffel coat	magic carpet
bow tie	elastic band	magnetic tape
calendar month	end result	main course
camp follower	estate agent	male chauvinist
canon law	eye candy	marital status
carbon copy	fairy tale	market share
card index	fallow deer	message board
carriage clock	false friend	milk tooth
cat's cradle	false start	motion sickness
centre forward	fine art	mountain bike
charge account	fire alarm	negative equity
chick flick	fire hydrant	night owl
chicken wire	first class	number one
child abuse	fitting room	open letter
chimney sweep	flood plain	orchestra pit
cleft palate	fruit machine	own goal

Table A.1 (*cont.*)

paper tiger	rope ladder	systems analyst
party line	safety pin	table tennis
paving stone	safety razor	tape measure
personal ad	safety valve	tax avoidance
petrol station	salad bar	taxi rank
place mat	sandwich course	time zone
poetic licence	sea lane	toggle switch
police dog	security guard	top gear
polling booth	security risk	trade union
pop art	senior citizen	treasure hunt
postage stamp	service charge	tumble dryer
pot plant	service industry	turning circle
powder room	sheet lightning	turning point
power station	shop steward	vacant possession
precious stone	shoulder strap	vending machine
prickly pear	sixth sense	video camera
private practice	ski jump	voice box
public house	slow motion	wagon train
quad bike	slumber party	walking stick
race meeting	snake charmer	water biscuit
racing car	sod all	whipping boy
rain check	sour cream	white goods
reception room	space station	white trash
red card	spirit level	wide boy
registry office	spot check	wild boar
rhythm section	square root	wooden spoon
right angle	stomach pump	work permit
rock pool	storm cloud	world power
rock salt	straw poll	wrapping paper
rocket science	subordinate clause	youth hostel
roller blind	suspender belt	zip code
room temperature	swimming pool	

Table A.2 *The 200 OHS_600 compounds with exclusively hyphenated spelling*

able-bodied	big-hearted	candy-striped
all-important	blue-blooded	cash-strapped
all-round	booze-up	chain-smoke
all-star	bottom-up	clean-shaven
also-ran	broken-down	clear-cut
bad-tempered	bust-up	clear-sighted
balls-up	by-product	close-fitting
bell-bottoms	camera-shy	close-set

Table A.2 (*cont.*)

close-up	hands-off	make-up
colour-blind	hard-nosed	mind-blowing
conscience-stricken	hard-wired	mind-boggling
cost-effective	hard-working	mix-up
creepy-crawly	has-been	modern-day
cure-all	hawk-eyed	moth-eaten
deep-set	heat-seeking	nail-biting
double-check	he-man	name-calling
double-edged	high-end	narrow-minded
double-jointed	high-flown	new-found
drip-dry	high-level	nice-looking
drive-in	high-pitched	no-frills
ear-splitting	high-pressure	no-nonsense
earth-shattering	high-risk	off-centre
even-tempered	high-spirited	off-peak
eye-opener	hold-up	once-over
face-saving	hot-blooded	one-sided
fact-finding	hot-wire	one-stop
fair-minded	house-sit	one-way
far-fetched	house-trained	on-screen
far-off	ice-cold	open-mouthed
far-reaching	ill-assorted	open-plan
feeble-minded	ill-fitting	part-time
fifty-fifty	ill-mannered	peace-loving
first-hand	ill-starred	pie-eyed
first-rate	in-between	pin-up
flat-footed	industrial-strength	pistol-whip
frame-up	in-store	punch-up
free-range	jam-packed	purpose-built
freeze-frame	jumped-up	quick-witted
fry-up	kind-hearted	radio-controlled
full-bodied	knees-up	rake-off
full-grown	know-all	red-faced
full-length	labour-saving	red-hot
full-page	lay-by	right-hand
get-together	lead-in	right-handed
glow-worm	light-fingered	roll-on
go-ahead	light-headed	second-guess
go-between	like-minded	second-rate
go-getter	long-lived	shoo-in
goggle-eyed	long-term	short-change
go-slow	low-fat	short-lived
grant-maintained	low-key	short-range
grown-up	low-pitched	simple-minded
gut-wrenching	low-rent	sit-in
habit-forming	low-rise	six-shooter

Table A.2 (*cont.*)

sky-high	strong-willed	wafer-thin
small-scale	third-rate	warm-blooded
small-town	thought-provoking	warm-hearted
snarl-up	to-do	washing-up
snow-white	tongue-tied	water-repellent
so-called	top-heavy	weak-willed
soft-boiled	top-level	weather-beaten
sol-fa	touch-type	well-nigh
soul-searching	two-bit	white-hot
spoon-feed	two-edged	win-win
stand-up	unlooked-for	world-beater
stop-go	user-friendly	would-be
strong-minded	vacuum-packed	

Table A.3 *The 200 OHS_600 compounds with exclusively solid spelling*

adman	brownstone	firearm
aircraft	brushwood	fireball
aircrew	candyfloss	firefly
airport	carsick	fireguard
airway	carthorse	firewall
arsehole	checkpoint	flagstone
ashtray	churchman	footbridge
backside	coalfield	footstool
barbershop	coastline	foxhound
bareback	cockpit	freeman
basketball	cookout	freeway
battledress	crowbar	frogman
beachhead	daylight	gaolbird
bedroom	deadhead	gasbag
birthplace	dogfight	gatehouse
birthright	doughnut	goalmouth
blacklist	dragonfly	godson
blacksmith	drumbeat	gravestone
boathouse	drumstick	greyhound
bonehead	earache	gridlock
borderline	earpiece	grindstone
borehole	earthbound	groundsheet
breakneck	eggshell	groundswell
breastbone	evensong	hairstyle
bridegroom	eyebrow	hallway
bridgehead	eyeglass	handbook
broadsheet	farmhouse	handrail

Table A.3 (*cont.*)

handset	payload	sunrise
hangdog	peephole	sunstroke
hardball	penknife	sweatshirt
headache	playground	swimsuit
heartburn	plaything	switchblade
heartland	poorhouse	swordfish
hemline	quarterdeck	teamwork
hoodwink	racehorse	tenderloin
horseman	railroad	textbook
hothead	ripcord	threadbare
housewife	rustproof	thundercloud
inbred	sandalwood	timepiece
indeed	sawmill	timeshare
kettledrum	schoolgirl	tinderbox
keypad	scoreline	toolbar
knucklehead	sealskin	towpath
lampshade	searchlight	tripwire
landfill	seashore	tumbleweed
lifeline	seaweed	turnstile
livestock	shellfish	watercross
lookout	shoreline	waterfront
mainstream	shorthand	watermark
markdown	sickbed	weatherman
matchstick	skinhead	weekday
mayfly	skylark	weekend
meanwhile	skylight	westbound
middleman	snakebite	wheelbase
milkman	snowdrop	whiteboard
moonlit	songbook	wholemeal
motherboard	southbound	wildfire
motorboat	spacecraft	windfall
mouthpiece	spacesuit	windowsill
muscleman	stagehand	wingspan
nearby	standpoint	womankind
nightcap	standstill	woodshed
numbskull	starfish	woodwind
nursemaid	steamship	workday
oatmeal	stopcock	workforce
pasteboard	storehouse	workweek
patrolman	stronghold	

Table A.4 *The ninety-five biconstituent CompText compounds with open spelling*

added value	handing back	prison staff
adrenaline rush	hard drug	red cabbage
all right	homeless hostel	red carpet
ankle tag	human being	red onion
arrival time	immigration form	sailing time
average age	imperial power	sale price
back exit	jail term	scale model
best value	jalapeno chilli	settler community
bicycle commute	language guru	short fuse
bike ride	legal team	sister ship
bin bag	lime juice	slave wage
business community	lip liner	slider attachment
cabinet minister	luggage tag	social theorist
chicken breast	manila rope	special adviser
commodity chain	marital coercion	speed peeler
company badge	material culture	spring afternoon
copper alloy	media camera	star struck
costs hearing	membership term	star treatment
cultural theorist	military funeral	station platform
customs guide	minimum term	styling secret
departure blast	national hero	styling suite
drinking club	northern sea	survival hope
economic theory	nutritional information	three thousand
education secretary	ocean liner	timed cuffew
electronic tag	ocean romance	top flight
fellow resident	odd job	Tuesday afternoon
five thousand	on hand	TV elite
font door	open jail	value theory
four thousand	pork chop	white cabbage
free header	PR adviser	wine suggestion
hair hero	press photographer	yes vote
hair partner	prime minister	

Table A.5 *The twenty-one biconstituent CompText compounds with hyphenated spelling*

all-over	ever-growing	new-age
battle-zone	iron-bound	north-west
blow-dry	last-minute	sailor-speak
bottom-line	long-locked	ship-crowded
cross-legged	long-unopened	snarled-up
eight-month	low-income	steam-horn
empire-builder	never-exhibited	T-shirt

Table A.6 *The thirty-three biconstituent CompText compounds with solid spelling*

alongside	everyone	perhaps
Another	everything	something
Anybody	fisherman	sometimes
Anyone	grandmother	somewhere
Anything	however	Sunday
Anywhere	indoors	thereby
Bridgewing	lifeworld	tidestream
Craftsman	maybe	whatever
Crossroad	newspaper	within
Everybody	nothing	without
Everyday	onside	without

Table A.7 *English lexical suffixes and final combining forms (adapted from Sanchez 2008: 137–139)*

Suffix	Example
-able	<i>acceptable</i>
-age	<i>package</i>
-al	<i>conventional</i>
-ally	<i>specifically</i>
-ance	<i>performance</i>
-ant	<i>servant</i>
-ar	<i>familiar</i>
-ary	<i>parliamentary</i>
-ate	<i>operate</i>
-ation	<i>variation</i>
-ative	<i>conservative</i>
-cy	<i>efficiency</i>
-dom	<i>freedom</i>
-ed	<i>armed</i>
-ee	<i>employee</i>
-en	<i>threaten</i>
-ence	<i>difference</i>
-ency	<i>emergency</i>
-ent	<i>ancient</i>
-er	<i>teacher</i>
-ery	<i>gallery</i>
-ette	<i>cigarette</i>
-free	<i>salt-free</i>
-ful	<i>powerful</i>

Table A.7 (*cont.*)

Suffix	Example
-fy	<i>satisfy</i>
-ial	<i>commercial</i>
-ian	<i>politician</i>
-ible	<i>terrible</i>
-ic	<i>democratic</i>
-ical	<i>historical</i>
-ice	<i>justice</i>
-iety	<i>variety</i>
-ify	<i>identify</i>
-ing	<i>meeting</i>
-ion	<i>suggestion</i>
-ise	<i>realise</i>
-ism	<i>mechanism</i>
-ist	<i>artist</i>
-ite	<i>opposite</i>
-ition	<i>definition</i>
-itude	<i>attitude</i>
-ity	<i>community</i>
-ive	<i>effective</i>
-less	<i>helpless</i>
-like	<i>ladylike</i>
-ly	<i>friendly</i>
-ment	<i>movement</i>
-ness	<i>awareness</i>
-nomy	<i>economy</i>
-ology	<i>technology</i>
-or	<i>mirror</i>
-ory	<i>statutory</i>
-our	<i>behaviour</i>
-ous	<i>famous</i>
-ry	<i>ministry</i>
-scape	<i>landscape</i>
-self	<i>herself</i>
-ship	<i>leadership</i>
-sion	<i>extension</i>
-t	<i>complaint</i>
-th	<i>growth</i>
-tion	<i>intervention</i>
-ty	<i>beauty</i>
-ual	<i>sexual</i>
-ure	<i>pressure</i>
-ward	<i>forward</i>
-wards	<i>afterwards</i>
-wise	<i>otherwise</i>
-y	<i>assembly</i>

Table A.8 *English prefixes and initial combining forms (adapted from Sanchez 2008: 136)*

Prefix	Example
ad-	<i>adoral</i>
after-	<i>afternoon</i>
agri-	<i>agriculture</i>
back-	<i>backyard</i>
centi-	<i>per cent</i>
circum-	<i>circumstance</i>
co-	<i>colleague</i>
com-	<i>combine</i>
con-	<i>contemporary</i>
de-	<i>decline</i>
deca-	<i>decade</i>
di-	<i>divide</i>
dis-	<i>disappear</i>
en-	<i>enable</i>
im-	<i>impossible</i>
in-	<i>independent</i>
inter-	<i>international</i>
intro-	<i>introduce</i>
mid-	<i>mid-sentence</i>
mini-	<i>minor</i>
mis-	<i>mistake</i>
out-	<i>outside</i>
over-	<i>overcome</i>
phono-	<i>telephone</i>
post-	<i>post-war</i>
pre-	<i>predict</i>
re-	<i>recall</i>
sub-	<i>subsequent</i>
super-	<i>supreme</i>
tele-	<i>telephone</i>
trans-	<i>transfer</i>
un-	<i>unable</i>
uni-	<i>unique</i>
wh-	<i>who</i>
with-	<i>withdraw</i>

Table A.9 Significant variables for OHS_600 compound spelling

	Variable	Explanation	Value	O	H	S	Preferred spellings
A1	o_1_CC_2_r	Consonant cluster across boundary	4–6	–	+	+	2
A2	Ident_lett_r	Repeated letters across boundary	+	+	+	--	2
A4	o_1_VV_2_r	Vowels across boundary	+	0	++	--	I+
B2a	Syll_total_r	Number of syllables	2	–	0	+	I
B2b	Syll_total_r	Number of syllables	3	+	0	–	I
B2c	Syll_total_r	Number of syllables	4 or more	++	0	--	I+
B3a	Lett_total_r	Number of letters	4–8	–	+	++	I
B3b	Lett_total_r	Number of letters	11 or more	++	0	--	I+
B5	Lett_diff_12	Length difference between constituents exceeds 1:2 (letters)	+	–	++	--	I+
B6	Syll_diff_12	Length difference between constituents exceeds 1:2 (syllables)	+	++	0	--	I+
B7a	Lett_1_r	Length of first constituent (letters)	2	--	++	–	I+
B7b	Lett_1_r	Length of first constituent (letters)	7 or more	++	–	–	I+
B7c	Lett_2_r	Length of second constituent (letters)	2	--	++	--	I+
B7d	Lett_2_r	Length of second constituent (letters)	3–5	0	–	+	I

Table A.9 (cont.)

	Variable	Explanation	Value	O	H	S	Preferred spellings
B7e	Lett_2_r	Length of second constituent (letters)	6 or more	+	+	--	2
B8a	Syll_1_r	Length of first constituent (syllables)	1	-	+	+	2
B8b	Syll_1_r	Length of first constituent (syllables)	2 or more	++	-	--	I+
B8c	Syll_2_r	Length of second constituent (syllables)	1	-	0	+	I
B8d	Syll_2_r	Length of second constituent (syllables)	2	+	+	--	2
B8e	Syll_2_r	Length of second constituent (syllables)	3 or more	++	0	--	I+
C1a	Total_BNC_r	Frequency	high (127-55,000)	-	+	+	2
C1b	Total_BNC_r	Frequency	low (0-17)	+	0	-	I
C2	Freq_rvs2	Combined frequency ranges of the first and second constituents	low+low (0-1,516 + 0-832)	0	+	-	I
C3a	Freq_rvs2	Combined frequency ranges of the first and second constituents	high+high (21,569-2,360,010 + 20,618-3,530,089)	-	+	-	I
C4a	Freq_diff_r2	Frequency difference between the constituents	high (1:50/50:1-1:99/99:1)	0	+	-	I
C4b	Freq_diff_r2	Frequency difference between the constituents	very high (exceeding 1:100/100:1)	-	++	-	I+
C4Aa	Freq_1_r	Frequency range of the first constituent	low	+	-	-	I
C4Ab	Freq_1_r		high	-	+	-	I

Frequency range of the first constituent

C4Ba	Freq_2_r	Frequency range of the second constituent	low (0–832)	–	+	–	I
C4Bb	Freq_2_r	Frequency range of the second constituent	high	–	+	+	2
D2	Stress	Main stress	on the first constituent	0	–	+	I
D3	Stress	Main stress	on the second constituent	+	++	–	I
D4a	Stress	Main stress	on the second constituent	++	–	–	I+
			+				
			in noun compounds				
D4b	Stress	Main stress	on the first constituent	0	–	+	I
			+				
			in noun compounds				
E1	Nonfin_lex_suff	Non-compound-final lexical suffix	+	++	–	–	I+
E4	Final_ingeder_r	Compound-final suffix <i>-ing, -ed</i> or <i>-er</i>	+	–	++	–	I+
E6a	Complex_const_r	One or more complex constituents	+	+	++	–	I
E6b	Complex_const_r	One or more complex constituents	–	0	–	+	I
E7a	Morphol_struct_r	Head-final morphological structure	+	+	–	0	I
E7b	Morphol_struct_r	Head-final morphological structure	–	–	++	0	I+

Table A.9 (cont.)

Variable	Explanation	Value	O	H	S	Preferred spellings
F1	PoS_comp_r	n	+	-	+	2
F2	PoS_comp_r	adj	--	++	-	I+
F6a	PoS_1	n	+	-	+	2
F6b	PoS_1	adj	+	0	-	I
F6c	PoS_1	v	--	++	+	I
F6d	PoS_1	adv	--	++	--	I+
F6e	PoS_1	g	--	++	--	I+
F7a	PoS_2_r	n	+	-	+	2
F7b	PoS_2_r	adj	--	++	-	I+
F7c	PoS_2_r	g	--	++	--	I+
F7d	PoS_2_r	v	--	++	--	I+
F7e	PoS_2_r	adv	--	++	--	I+
F8	Lexgr_12_diff	+	--	++	--	I+

G4a	Idiom	Idiomatcity	+	(i)	-	+	+	2
G4b	Idiom	Idiomatcity	+	(h)	-	+	o	I
G4c	Idiom	Idiomatcity	+	(l)	+	-	o	I
H1	Mixed_etym	Germanic+Romance constituents	+	+	+	o	-	I
H2	Foreignness	Synchronically felt foreignness	+	+	++	-	-	I+
H3	Age_r	Age of the compound (= first OED attestation)	-	-	--	-	++	I+
H4	Age_r	Age of the compound (= first OED attestation)	++	+	+	+	-	I
H5a	Earl_spell_r	Earliest attested spelling in the OED	+	o	+	-	o	I
H5b	Earl_spell_r	Earliest attested spelling in the OED	-	h	-	+	o	I
H5c	Earl_spell_r	Earliest attested spelling in the OED	--	s	--	-	++	I+
J1a	LS_r	Spelling preferred by left constituent family size	+	o	+	-	-	I
J1b	LS_r	Spelling preferred by left constituent family size	--	h	--	++	-	I+
J1c	LS_r	Spelling preferred by left constituent family size	-	s	-	o	+	I
J2a	RS_r	Spelling preferred by right constituent family size	++	o	++	-	-	I+
J2b	RS_r	Spelling preferred by right constituent family size	--	h	--	++	--	I+
J2c	RS_r	Spelling preferred by right constituent family size	-	s	-	-	+	I

Table A.9 (*cont.*)

	Variable	Explanation	Value	O	H	S	Preferred spellings
J3a	LF_r	Spelling preferred by left constituent family frequency	o	+	-	-	I
J3b	LF_r	Spelling preferred by left constituent family frequency	h	-	+	o	I
J3c	LF_r	Spelling preferred by left constituent family frequency	s	-	o	+	I
J4a	RF_r	Spelling preferred by right constituent family frequency	o	+	-	-	I
J4b	RF_r	Spelling preferred by right constituent family frequency	h	--	++	--	I+
J4c	RF_r	Spelling preferred by right constituent family frequency	s	-	-	+	I

Table A.10 *Spelling tendencies without statistical backing in OHS_600*

•	Garden path clusters at the constituent boundaries favour solid spelling. [A3; significant in OHS_extra]
•	Capitalised non-initial constituents favour open spelling and disfavour solid spelling. [A5]
•	A compound-medial apostrophe favours open spelling and disfavors solid spelling. [A6; significant in OHS_extra]
•	Three or more constituents disfavour solid spelling. [B1]
•	A single-letter constituent disfavors solid spelling. [B4]
•	Non-compound-final grammatical suffixes disfavour solid spelling. [E2]
•	Non-compound-final grammatical suffixes favour hyphenation. [E2]
•	Non-compound-initial prefixes disfavour solid spelling. [E3]
•	Non-compound-initial prefixes favour open spelling. [E3]
•	Hyphenated constituents (prefixations or combining forms) favour open spelling. [E5]
•	Acronymic constituents favour open spelling. [E6]
•	Verb compounds disfavour open spelling. [F4]
•	Verb compounds consisting of two verbs favour hyphenation. [F4]
•	Adverb compounds disfavour open spelling. [F5]
•	Attributive position favours hyphenation of open compounds. [F9]
•	Predicative position favours open spelling of hyphenated compounds. [F10]
•	A compound-final general reference noun favours solid spelling. [G1]
•	Species-genus compounds favour solid spelling. [G2]
•	Compounds with identical constituents favour hyphenation. [G3]
•	Co-hyponymous constituents favour solid spelling. [G3]
•	Meronymous constituents the first of which is the larger entity favour solid spelling. [G3]
•	The combination of one Germanic and one Romance constituent disfavors solid spelling. [H1]

Table A.11 *Grammatical compounds and their spelling in the Oxford English Dictionary*

Compound	PoS	Spelling
<i>anyhow</i>	conj	s
<i>anyway</i>	conj	s
<i>anyways</i>	conj	s
<i>whenever</i>	conj	s
<i>whereas</i>	conj	s
<i>whereup</i>	conj	s
<i>howbeit</i>	conj	ss
<i>notwithstanding</i>	conj	ss
<i>sobeit</i>	conj	ss
<i>whencesoever</i>	conj	ss

Table A.II (cont.)

Compound	PoS	Spelling
<i>whensoever</i>	conj	ss
<i>wheresoever</i>	conj	ss
<i>whethersoever</i>	conj	ss
<i>boom-boom</i>	int	h
<i>easy-peasy</i>	int	h
<i>fiddle-faddle</i>	int	h
<i>gee-ho</i>	int	h
<i>goo-goo</i>	int	h
<i>haw-haw</i>	int	h
<i>heigh-ho</i>	int	h
<i>ho-hum</i>	int	h
<i>huh-uh</i>	int	h
<i>night-night</i>	int	h
<i>rap-tap</i>	int	h
<i>weet-weet</i>	int	h
<i>wham-bam</i>	int	h
<i>wham-bang</i>	int	h
<i>whizz-bang</i>	int	h
<i>who-whoop (who-whoop)</i>	int	h
<i>ta-ta (tata, ta ta)</i>	int	h (s, o)
<i>hey-ho, hey ho</i>	int	h, o
<i>hey-day, heyday</i>	int	h, s
<i>yo-heave-ho</i>	int	hh
<i>gee whizz</i>	int	o
<i>ha ha</i>	int	o
<i>heads up</i>	int	o
<i>heave ho</i>	int	o
<i>hot damn</i>	int	o
<i>lovely jubbly</i>	int	o
<i>oh yeah</i>	int	o
<i>oh, oh (ohoh)</i>	int	o (s)
<i>pooh pooh</i>	int	o
<i>right on</i>	int	o
<i>what ho</i>	int	o
<i>down and dirty</i>	int	oo
<i>checkmate</i>	int	s
<i>farewell</i>	int	s
<i>half-way, halfway</i>	prep	h, s
<i>out of</i>	prep	o
<i>round about</i>	prep	o
<i>near by (near-by, nearby)</i>	prep	o (h, s)
<i>on to, onto</i>	prep	o, s
<i>to and fro</i>	prep	oo
<i>up and down</i>	prep	oo

Table A.II (cont.)

Compound	PoS	Spelling
<i>alongside</i>	prep	s
<i>endlong</i>	prep	s
<i>inboard</i>	prep	s
<i>inside</i>	prep	s
<i>into</i>	prep	s
<i>midway</i>	prep	s
<i>outside</i>	prep	s
<i>throughout</i>	prep	s
<i>upon</i>	prep	s
<i>within</i>	prep	s
<i>without</i>	prep	s
<i>notwithstanding</i>	prep	ss
<i>withoutside</i>	prep	ss
<i>such-like, suchlike</i>	pron	h, s
<i>some one, someone</i>	pron	o, s
<i>another</i>	pron	s
<i>anybody</i>	pron	s
<i>anything</i>	pron	s
<i>anywhat</i>	pron	s
<i>everybody</i>	pron	s
<i>everything</i>	pron	s
<i>whatever</i>	pron	s
<i>whichever</i>	pron	s
<i>whoever</i>	pron	s
<i>whomever</i>	pron	s
<i>whatsoever</i>	pron	ss
<i>whethersoever</i>	pron	ss
<i>whomsoever</i>	pron	ss
<i>whosoever</i>	pron	ss

Note that the comma is part of the compound interjection *oh, oh*. Comma-separated lemmatised alternatives from the OED are considered of equal importance and separated by a comma in Table A.II. Secondary alternatives, which are introduced by *Also* in the OED, are enclosed by parentheses in Table A.II. Note that *somebody* is missing in the list because it is classified as a noun in the OED. *Nobody* has no part-of-speech indication, and the following grammatical items from the Master List are not lemmatised with part of speech in the OED either:

- the pronouns *each+other*, *no+one*, *one+another* and *such+and+such*
- the prepositions *according+to*, *apart+from*, *due+to*, *next+to*, *owing+to* and *relating+to*
- the conjunction *in+as+much+as*
- the interjections *aw+shucks*, *full+stop*, *oops+a+daisy* and *thank+you*.

Since one might expect to find additional grammatical compounds in other dictionaries, the OED on its own cannot be considered exhaustive with regard to the listing of grammatical compounds.

References

- Aarts, Bas. 2011. *Oxford modern grammar*. Oxford: Oxford University Press.
- Achtert, Walter S. 1985. *The MLA style manual*. New York: Modern Language Association.
- Adams, Valerie. 1973. *An introduction to modern English word-formation*. London: Longman.
- Adams, Valerie. 2001. *Complex words in English*. Harlow: Pearson.
- Aitchison, Jean. 1991. *Language change: Progress or decay?* 2nd edn. Cambridge: Cambridge University Press.
- Aitchison, Jean. 1994. *Words in the mind: An introduction to the mental lexicon*. 2nd edn. Oxford: Blackwell.
- Allan, Keith and Kate Burridge. 2006. *Forbidden words: Taboo and the censoring of language*. Cambridge: Cambridge University Press.
- Antos, Gerd. 2003. "Imperfektibles" sprachliches Wissen: Theoretische Vorüberlegungen zu "sprachlichen Zweifelsfällen". *Linguistik Online* 16(4). 35–46.
- Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.
- Aston, Guy and Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- Augst, Gerhard. 2005. The German spelling reform: An example for the Simplified Spelling Society. In Bernadette Hughes (ed.), *Spelcon 2005 conference report*. 21–29.
- Baayen, Harald R. 2012. Demythologizing the word frequency effect. In Gary Libben, Gonia Jarema and Chris Westbury (eds.), *Methodological and analytic frontiers in lexical research*. Amsterdam: Benjamins. 171–195.
- Bailey, Charles-James N. 1979. *English punctuation in brief*. Berlin: Technische Universität.
- Bartsch, Sabine, Stefan Evert, Thomas Proisl and Peter Uhrig. 2015. (Association) measure for measure: Comparing collocation dictionaries with co-occurrence data for a better understanding of the notion of collocation. Paper presented at ICAME 36, Trier.
- Barz, Irmhild. 1993. Graphische Varianten bei der substantivischen Komposition. *Deutsch als Fremdsprache: Zeitschrift zur Theorie und Praxis des Deutschunterrichts für Ausländer* 30(3). 167–172.

- Bauer, Laurie. 1978. *The grammar of nominal compounding with special reference to Danish, English and French*. Odense: Odense University Press.
- Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.
- Bauer, Laurie. 1994. *Watching English change: An introduction to the study of linguistic change in standard Englishes in the twentieth century*. London: Longman.
- Bauer, Laurie. 1998. When is a sequence of two nouns a compound in English? *English Language and Linguistics* 2(1). 65–86.
- Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.
- Bauer, Laurie. 2003. *Introducing linguistic morphology*. 2nd edn. Edinburgh: Edinburgh University Press.
- Bauer, Laurie. 2005. The borderline between derivation and compounding: Selected papers from the 11th Morphology Meeting, Vienna, February 2004. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer and Franz Rainer (eds.), *Morphology and its demarcations*. Amsterdam: Benjamins. 97–108.
- Bauer, Laurie. 2009. Typology of compounds. In Rochelle Lieber and Pavol Štekauer (eds.), *The Oxford handbook of compounding*. Oxford: Oxford University Press. 343–356.
- Beal, Joan. 2010. The grocer's apostrophe: Popular prescriptivism in the 21st century. *English Today* 26(2). 57–64.
- Beal, Joan. 2012. New authorities and the 'New Prescriptivism'. In Anne Schröder, Ulrich Busse and Ralf Schneider (eds.), *Codification, canons and curricula: Description and prescription in language and literature*. Bielefeld: Aisthesis. 183–193.
- Behrens, Heike. 2007. The acquisition of argument structure. In Thomas Herbst and Katrin Götz-Votteler (eds.), *Valency: Theoretical, descriptive and cognitive issues*. Berlin: De Gruyter. 193–214.
- Bell, Masha. 2009. *Rules and exceptions of English spelling*. Cambridge: Pegasus.
- Bell, Melanie and Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. *Journal of Linguistics* 48(3). 485–520.
- Bertram, Raymond, Victor Kuperman, R. Harald Baayen and Jukka Hyönä. 2011. The hyphen as a segmentation cue in compound processing: It's getting better all the time. *Scandinavian Journal of Psychology* 52(6). 530–544.
- Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.
- Bieswanger, Markus. 2007. 2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space- and time-saving strategies in English and German text messages. In Simeon Floyd, Taryne Hallet, Sae Oshima and Aaron Shield

- (eds.), *Texas Linguistic Forum* 50. <http://studentorgs.utexas.edu/salsa/proceedings/2006/Bieswanger.pdf>. (25 June 2011.)
- Bolinger, Dwight L. 1967. Adjectives in English: Attribution and predication. *Lingua* 18(1). 1–34.
- Bollée, Annegret. 1994/95. *Die französische Orthographie*. Bamberg: script.
- Booij, Geert. 2007. *The grammar of words: An introduction to morphology*. 2nd edn. Oxford: Oxford University Press.
- Booth, Wayne C., Gregory G. Colomb and Joseph M. Williams. 2008. *The craft of research*. 3rd edn. Chicago: University of Chicago Press.
- Bredel, Ursula. 2011. *Interpunktion*. Heidelberg: Winter.
- Brinton, Laurel J. and Leslie K. Arnovick. 2006. *The English language: A linguistic history*. Oxford: Oxford University Press.
- Burgschmidt, Ernst. 1973. System und Norm im Bereich der Wortbildung. In Ernst Burgschmidt (ed.), *System, Norm und Produktivität in der Wortbildung*. Vol. I. Erlangen: Seminar für Englische Philologie. 1–172.
- Burgschmidt, Ernst. 1977. Strukturierung, Norm und Produktivität in der Wortbildung. In Herbert Ernst Brekle and Dieter Kastovsky (eds.), *Perspektiven der Wortbildungsforschung*. Bonn: Bouvier. 39–47.
- Burridge, Kate. 2005. *Weeds in the garden of words: Further observations on the tangled history of the English language*. Cambridge: Cambridge University Press.
- Burridge, Kate. 2010. Linguistic cleanliness is next to godliness: Taboo and purism. *English Today* 26(2). 3–13.
- Busse, Ulrich and Anne Schröder. 2010a. How Fowler became ‘The Fowler’: An anatomy of a success story. *English Today* 26(2). 45–54.
- Busse, Ulrich and Anne Schröder. 2010b. Problem areas of English grammar between usage, norm, and variation. In Alexandra N. Lenz and Albrecht Plewnia (eds.), *Grammar between norm and variation*. Frankfurt am Main: Peter Lang. 87–102.
- Bußmann, Hadumod. 2002. *Lexikon der Sprachwissenschaft*. 3rd edn. Stuttgart: Kröner.
- Butcher, Judith. 1992. *Copy-editing: The Cambridge handbook for editors, authors and publishers*. 3rd edn. Cambridge: Cambridge University Press.
- Bybee, Joan. 1998. The emergent lexicon. *Chicago Linguistic Society* 34. 421–435.
- Bybee, Joan. 2003. *Phonology and language use*. 2nd edn. Cambridge: Cambridge University Press.
- Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Cambridge advanced learner's dictionary* (CALD). 2008. 3rd edn. with CD-ROM. Cambridge: Cambridge University Press.
- Cappelle, Bert. 2009. Can we factor out free choice? In Andreas Dufter, Jürg Fleischer and Guido Seiler (eds.), *Describing and modeling variation in grammar*. Berlin: Mouton de Gruyter. 183–201.
- Carey, Gordon Vero. 1957. *Punctuation*. Cambridge: Cambridge University Press.

- Carey, Gordon Vero. 1958. *Mind the stop: A brief guide to punctuation with a note on proof-correction*. New edn. Cambridge: Cambridge University Press.
- Carney, Edward. 1994. *A survey of English spelling*. London: Routledge.
- Carney, Edward. 1997. *English spelling*. London: Routledge.
- Celce-Murcia, Marianne, Donna M. Brinton and Janet M. Goodwin. 1996. *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge: Cambridge University Press.
- Chambers, Jack K. and Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.
- Clark, John Owen Edward. 1990. *Harrap's English punctuation & hyphenation*. London: Harrap.
- Coates, Jennifer. 2004. *Women, men and language: A sociolinguistic account of gender differences in language*. 3rd edn. Harlow: Pearson.
- Colff, Adri van der. 1998. Die Komposita in Langenscheidts Großwörterbuch Deutsch als Fremdsprache. In Herbert Ernst Wiegand (ed.), *Perspektiven der pädagogischen Lexikographie des Deutschen: Untersuchungen anhand von "Langenscheidts Großwörterbuch Deutsch als Fremdsprache"*. Tübingen: Niemeyer. 193–207.
- Collins English dictionary* (CED). 2004. 6th edn. London: HarperCollins.
- Coseriu, Eugenio. 1978. *Probleme der strukturellen Semantik*. Tübingen: Narr.
- Croft, William. 2002. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press.
- Croft, William and David Alan Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.
- Cruse, David Alan. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.
- Cruse, David Alan. 1990. Prototype theory and lexical semantics. In Savas L. Tsohatzidis (ed.), *Meanings and prototypes: Studies in linguistic categorization*. London: Routledge. 382–402.
- Crystal, David. 1997. *The Cambridge encyclopedia of language*. 2nd edn. Cambridge: Cambridge University Press.
- Crystal, David. 2001. *Language and the internet*. Cambridge: Cambridge University Press.
- Cullen, Kay. 1999. *Chambers guide to punctuation*. Edinburgh: Chambers Harrap.
- Czerlinski, Jean, Gerd Gigerenzer and Daniel G. Goldstein. 1999. How good are simple heuristics? In Gerd Gigerenzer, Peter M. Todd and the ABC Research Group (eds.), *Simple heuristics that make us smart*. Oxford: Oxford University Press. 97–118.
- Dąbrowska, Ewa. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2 (3). 219–253.
- Daelemans, Walter, Jakub Zavrel, Ko van der Sloot and Anton van den Bosch. 2004. *TiMBL: Tilburg Memory Based Learner*, version 5.1, Reference Guide. ILK Technical Report 04 02, available from <http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf>.

- DeCapua, Andrea. 2008. *Grammar for teachers: A guide to American English for native and non-native speakers*. New York: Springer.
- Donalies, Elke. 2003. *Hochzeitstorte, laskaparasol, elmas küpe, cow's milk, casa de campo, cigarette-filtre, ricasdueñas . . .*: Was ist eigentlich ein Kompositum? *Deutsche Sprache: Zeitschrift für Theorie, Praxis, Dokumentation* 31(1). 76–93.
- Dressler, Wolfgang U. 2005. Compound types. In Gary Libben and Gonia Jarema (eds.), *The representation and processing of compound words*. Oxford: Oxford University Press. 23–44.
- Duden. 2006. *Duden: Die deutsche Rechtschreibung*. 24th edn. Mannheim: Duden.
- Dürscheid, Christa. 2011. Zweifeln als Chance? Zweifeln als Problem? Sprachliche Zweifelsfälle im Deutschunterricht. In Klaus-Michael Köpcke and Arne Ziegler (eds.), *Grammatik – Lehren, Lernen, Verstehen. Zugänge zur Grammatik des Gegenwartsdeutschen*. Berlin: de Gruyter. 155–173.
- Eckert, Hartwig and William Barry. 2005. *The phonetics and phonology of English pronunciation*. 2nd edn. Trier: Wissenschaftlicher Verlag Trier.
- Einhorn, Hillel J. 1982. Learning from experience and suboptimal rules in decision making. In Daniel Kahnemann, Paul Slovic and Amos Tversky (eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press. 268–283.
- Endley, Martin J. 2010. *Linguistic perspectives on English grammar: A guide for EFL teachers*. Charlotte, NC: Information Age Publishing.
- Epelboim, Julie, James R. Booth, T. Rebecca Ashkenazy, Arash Taleghani and Robert M. Steinmans. 1997. Fillers and spaces in text: The importance of word recognition during reading. *Vision Research* 37(20). 2899–2914.
- Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open-choice principle. *Text* 20(1). 29–62.
- Ernst, Jutta. 2000. Transatlantic simultaneity: John Dos Passos and Blaise Cendrars. In Klaus Martens (ed.), *Pioneering North America: Mediators of European culture and literature*. Würzburg: Königshausen & Neumann. 100–111.
- Faiß, Klaus. 1978. *Verdunkelte Compounds im Englischen*. Tübingen: Narr.
- Faiß, Klaus. 1981. Compound, pseudo-compound, and syntactic group especially in English. In Peter Kunsmann and Ortwin Kuhn (eds.), *Weltsprache Englisch in Forschung und Lehre: Festschrift für Kurt Wächtler*. Berlin: Schmidt. 132–150.
- Faiß, Klaus. 1992. *English historical morphology and word-formation: Loss versus enrichment*. Trier: Wissenschaftlicher Verlag Trier.
- Figueredo, Lauren and Connie K. Varnhagen. 2005. Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology: An International Quarterly* 26 (4–5). 441–458.
- Fiorentino, Robert and Ella Fund-Reznicek. 2009. Masked morphological priming of compound constituents. *The Mental Lexicon* 4(2). 159–193.

- Firth, John Rupert. 1957. A synopsis of linguistic theory, 1930–1955. In John Rupert Firth et al. (eds.), *Studies in linguistic analysis*. Oxford: Blackwell. 1–32.
- Fishman, Joshua A. 1977. Advances in the creation and revision of writing systems. In Joshua A. Fishman (ed.), *Advances in the creation and revision of writing systems*. The Hague: Mouton. XI–XXVIII.
- Foley, Mark and Diane Hall. 2003. *Longman advanced learners' grammar*. Harlow: Pearson.
- Forsyth, Eric N. and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*. 19–26.
- Fowler, Henry W. 1921. On hyphens. *Tract No. VI*. Oxford: Clarendon. 3–13.
- Fowler, Henry W. 1926. *A dictionary of modern English usage*. Oxford: Clarendon.
- Francis, W. Nelson and Henri Kučera. 1979. *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers*. Providence, RI: Department of Linguistics, Brown University.
- Frazier, Lyn. 1987. Sentence processing: A tutorial review. In Max Coltheart (ed.), *Attention and performance XII: The psychology of reading*. Hove: Lawrence Erlbaum. 559–586.
- Fries, Norbert. 2012. Spatien oder Die Bedeutung des Nichts. In Alexander Nebrig and Carlos Spoerhase (eds.), *Die Poesie der Zeichensetzung: Studien zur Stilistik der Interpunktion*. Berlin: Peter Lang. 407–428.
- Gallmann, Peter. 2004. Varianz in der Rechtschreibung. *Sprachspiegel* 2(1). 38–47.
- Ghomeshi, Jila. 2010. *Grammar matters: The social significance of how we use language*. Winnipeg: Arbeiter Ring Publishing.
- Ghomeshi, Jila, Ray Jackendoff, Nicole Rosen and Kevin Russell. 2004. Contrastive focus reduplication in English (the salad-salad paper). *Natural Language & Linguistic Theory* 22(2). 307–357.
- Giegerich, Heinz J. 2004. Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics* 8(1). 1–24.
- Gigerenzer, Gerd. 2007. *Gut feelings: The intelligence of the unconscious*. London: Penguin.
- Gigerenzer, Gerd. 2008. *Bauchentscheidungen: Die Intelligenz des Unbewussten und die Macht der Intuition*. 6th edn. Munich: Goldmann.
- Gigerenzer, Gerd and Daniel G. Goldstein. 1999. Betting on one good reason: The take the best heuristic. In Gerd Gigerenzer, Peter M. Todd and the ABC Research Group (eds.), *Simple heuristics that make us smart*. Oxford: Oxford University Press. 75–95.
- Gilquin, Gaëtanelle. 2006. The place of prototypicality in corpus linguistics: Causation in the hot seat. In Stefan Th. Gries and Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter. 159–191.
- Givón, Talmy. 1986. Prototypes: Between Plato and Wittgenstein. In Colette Craig (ed.), *Noun classes and categorization*. Amsterdam: Benjamins. 77–102.

- Goldberg, Adele. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Goldberg, Adele. 1996. Construction grammar. In Keith Brown and Jim Miller (eds.), *Concise encyclopedia of syntactic theories*. Oxford: Pergamon. 68–71.
- Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
- Goldstein, Norm. 2004. *Stylebook and briefing on media law*. New York: Basic Books.
- Görlach, Manfred. 1999. *English in nineteenth century England. An introduction*. Cambridge: Cambridge University Press.
- Götz, Dieter. 1971. *Studien zu den verdunkelten Komposita im Englischen*. Nuremberg: Hans Carl.
- GPO Style Manual: US Government Printing Office. 2008. *Style manual: An official guide to the form and style of federal government printing*. www.gpoaccess.gov/stylemanual/index.html. (4 July 2008.)
- Graddol, David. 1997. *The future of English?* London: British Council.
- Granger, Sylviane and Magali Paquot. 2008. Disentangling the phraseological web. In Sylviane Granger and Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam: Benjamins. 27–49.
- Green, Lee Albert and David R. Mehr. 1997. What alters physicians' decisions to admit to the coronary care unit? *The Journal of Family Practice* 45(3). 219–226.
- Greenbaum, Sidney. 1986. Spelling variants in British English. *Journal of English Linguistics* 19(2). 258–268.
- Grice, Paul. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and semantics. 3: Speech acts*. New York: Academic Press. 41–58.
- Gries, Stefan Th. 2003a. Grammatical variation in English: A question of 'structure vs. function'? In Günter Rohdenburg and Britta Mondorf (eds.), *Determinants of grammatical variation in English*. Berlin: de Gruyter. 155–173.
- Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1(1). 1–28.
- Gut, Ulrike. 2009. *Introduction to English phonetics and phonology*. Frankfurt am Main: Peter Lang.
- Guy, Gregory R. 2011. Variation and change. In Warren Maguire and April McMahon (eds.), *Analyzing variation in English*. Cambridge: Cambridge University Press. 178–198.
- Hacken, Pius ten. 1994. *Defining morphology: A principal approach to determining the boundaries of compounding, derivation and inflection*. Hildesheim: Olms.
- Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.
- Hale, Constance and Jessie Scanlon. 1999. *Wired style: Principles of English usage in the digital age*. New York: Broadway Books.
- Hall, Robert A. 1961. *Sound and spelling in English*. Philadelphia, PA: Chilton.
- Halliday, Michael, Alexander Kirkwood and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
- Handl, Susanne. 2008. Essential collocations for learners of English: The role of collocational direction and weight. In Fanny Meunier and Sylviane Granger

- (eds.), *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins. 43–66.
- Hanks, Patrick. 1988. Conventionality and efficiency in written English: The hyphen. *Journal of the Simplified Spelling Society* 2(1). 5–10.
- Hansen, Barbara, Klaus Hansen, Albrecht Neubert and Manfred Schentke. 1990. *Englische Lexikologie: Einführung in Wortbildung und lexikalische Semantik*. 3rd edn. Leipzig: VEB Verlag Enzyklopädie.
- Hart, Horace. 1957. *Rules for compositors and readers*. 37th edn. London: Oxford University Press.
- Hasher, Lynn and Rose T. Zacks. 1984. Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist* 39 (12). 1372–1388.
- Haspelmath, Martin. 2008. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–34.
- Hausmann, Franz Josef. 1985. Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In Henning Bergenholtz and Joachim Mugdan (eds.), *Lexikographie und Grammatik: Akten des Essener Kolloquiums zur Grammatik im Wörterbuch vom 28.-30.6.1984*. Tübingen: Niemeyer. 118–129.
- Hausmann, Franz Josef. 2004. Was sind eigentlich Kollokationen? In Kathrin Steyer (ed.), *Wortverbindungen – mehr oder weniger fest*. Berlin: de Gruyter. 309–334.
- Heisenberg, Werner. 1927. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik* 43(3). 172–198.
- Herbst, Thomas. 1996. What are collocations: Sandy beaches or false teeth. *English Studies* 77(4). 379–393.
- Herbst, Thomas. 2010. *English linguistics: A coursebook for students of English*. Berlin: de Gruyter.
- Herbst, Thomas. 2011. Choosing *sandy beaches*: Collocations, probabemes and the idiom principle. In Thomas Herbst, Susan Faulhaber and Peter Uhrig (eds.), *The phraseological view of language: A tribute to John Sinclair*. Berlin: de Gruyter. 27–57.
- Herbst, Thomas, David Heath, Ian F. Roe and Dieter Götz. 2004. *A valency dictionary of English*. Berlin: de Gruyter.
- Herbst, Thomas and Susan Schüller. 2008. *Introduction to syntactic analysis: A valency approach*. Tübingen: Narr.
- Hewings, Martin. 2005. *Advanced grammar in use*. 2nd edn. Cambridge: Cambridge University Press.
- Hillebrand, Ulrich. 1975. Chronologische und etymologische Untersuchungen zum französischen Wortbestand innerhalb der englischen Sprache. Münster: PhD thesis.
- Hoover, Regina M. 1971. Principles of English hyphenation. *College Composition and Communication* 22(2). 156–160.
- Horobin, Simon. 2013. *Does spelling matter?* Oxford: Oxford University Press.

- Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.
- Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Huddleston, Rodney and Geoffrey K. Pullum. 2012. *A student's introduction to English grammar*. Cambridge: Cambridge University Press.
- Hundt, Marianne, Andrea Sand and Rainer Siemund. 1998. *Manual of information to accompany the Freiburg – LOB Corpus of British English ('FLOB')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.
- Hundt, Marianne, Andrea Sand and Paul Skandera. 1999. *Manual of information to accompany the Freiburg – Brown Corpus of American English ('Frown')*. Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.
- Hundt, Markus. 2010. New norms: How new grammatical constructions emerge. In Alexandra N. Lenz and Albrecht Plewnia (eds.), *Grammar between norm and variation*. Frankfurt am Main: Peter Lang. 27–57.
- Inhoff, Albrecht Werner, Ralph Radach and Dieter Heller. 2000. Complex compounds in German: Interword spaces facilitate segmentation but hinder assignment of meaning. *Journal of Memory and Language* 42(1). 23–50.
- Jaarsveld, Henk J. van, Riet Coolen and Robert Schreuder. 1994. The role of analogy in the interpretation of novel compounds. *Journal of Psycholinguistic Research* 23(2). 111–137.
- Jackson, Howard and Etienne Zé Amvela. 2002. *Words, meaning and vocabulary: An introduction to modern English lexicology*. London: Continuum.
- Jacobs, Joachim. 2007. Vom (Un-)Sinn der Schreibvarianten. *Zeitschrift für Sprachwissenschaft* 26(special issue). 43–80.
- Jagemann, Hans C. G. von. 1900. Philology and purism. *PMLA (Journal of the Modern Language Association of America)* 15(1). 74–96.
- Johansson, Stig, Geoffrey N. Leech and Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo: Department of English, University of Oslo.
- Johnson, Sally. 2002. On the origin of linguistic norms: Orthography, ideology and the first constitutional challenge to the 1996 reform of German. *Language in Society* 31(4). 549–576.
- Johnson, Samuel. 1755. *A dictionary of the English language* [...]. 2 vols. London: Printed by W. Strahan for J. and P. Knapton etc.
- Jong, Nivja de, Laurie B. Feldman, Robert Schreuder, Matthew Pastizzo and R. Harald Baayen. 2002. The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language* 81 (1–3). 555–567.
- Juhász, Barbara J., Albrecht W. Inhoff and Keith Rayner. 2005. The role of interword spaces in the processing of English compound words. *Language and Cognitive Processes* 20(1–2). 291–316.

- Juhász, Barbara J., Matthew S. Starr, Albrecht W. Inhoff and Lars Placke. 2003. The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology* 94. 223–244.
- Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2). 137–194.
- Käge, Otmar. 1980. *Motivation: Probleme des persuasiven Sprachgebrauchs, der Metapher und des Wortspiels*. Göttingen: Kümmerle.
- Kahnemann, Daniel and Amos Tversky. 1982. Variants of uncertainty. In Daniel Kahnemann, Paul Slovic and Amos Tversky (eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press. 509–520.
- Kastovsky, Dieter. 2009. Diachronic perspectives. In Rochelle Lieber and Pavol Štekauer (eds.), *The Oxford handbook of compounding*. Oxford: Oxford University Press. 323–340.
- Kauhanen, Irina. 2006. Norms and sociolinguistic description. *SKY Journal of Linguistics* 19(special supplement). 34–46.
- Keller, Rudi. 1994. *Sprachwandel: Von der unsichtbaren Hand in der Sprache*. 2nd edn. Tübingen: Francke.
- Kiraz, George Anton and Bernd Möbius. 1998. Multilingual syllabification using weighted finite-state transducers. *3rd International Workshop on Speech Synthesis*. 71–76. www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_071.pdf. (18 May 2012)
- Kirby, Simon, Hannah Cornish and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *National Academy of Sciences of the United States of America (PNAS)* 105(31). 10681–10686.
- Klein, Wolf Peter. 2003. Sprachliche Zweifelsfälle als linguistischer Gegenstand: Zur Einführung in ein vergessenes Thema der Sprachwissenschaft. *Linguistik Online* 16(4). 5–33.
- Koch, Peter and Daniela Marzo. 2007. A two-dimensional approach to the study of motivation in lexical typology and its first application to French high-frequency vocabulary. *Studies in Language* 31 (2). 259–291.
- Kreiner, David S. and Philip B. Gough. 1990. Two ideas about spelling: Rules and word-specific memory. *Journal of Memory and Language* 29(1). 103–118.
- Kress, Gunther R. 2000. *Early spelling*. London: Routledge.
- Krott, Andrea, Harald Baayen and Robert Schreuder. 2001. Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics* 39(1). 51–93.
- Kruisinga, Etsko. 1932. *A handbook of present-day English. Part II. Accidence and syntax*. 5th edn. Groningen: P. Noordhoff. http://archive.org/stream/handbookofpresen22krui/handbookofpresen22krui_djvu.txt. (13 January 2014)
- Kuperman, Victor and Raymond Bertram. 2013. Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes* 28 (7). 939–966.

- Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Vol. 1. Theoretical prerequisites*. Stanford, CA: Stanford University Press.
- Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.
- Langenscheidt. *Taschenwörterbuch Englisch-Deutsch (TW)*. 2007. Munich: Langenscheidt.
- Langenscheidt/Collins. *Großwörterbuch Englisch-Deutsch*. 2008. Munich: Langenscheidt.
- Lass, Roger. 1997. *Historical linguistics and language change*. Cambridge: Cambridge University Press.
- Leech, Geoffrey. 1981. *Semantics: The study of meaning*. 2nd edn. London: Penguin.
- Leisi, Ernst and Christian Mair. 1999. *Das heutige Englisch: Wesenszüge und Probleme*. 8th, rev. edn. Heidelberg: Winter.
- Lenz, Alexandra N. and Albrecht Plewnia. 2010. On grammar between norm and variation. In Alexandra N. Lenz and Albrecht Plewnia (eds.), *Grammar between norm and variation*. Frankfurt am Main: Peter Lang. 7–25.
- Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.
- Lieber, Rochelle and Pavol Štekauer. 2009. Introduction. In Rochelle Lieber and Pavol Štekauer (eds.), *The Oxford handbook of compounding*. Oxford: Oxford University Press, 3–18.
- Lipka, Leonhard. 1977. Lexikalisierung, Idiomatisierung und Hypostasierung als Probleme einer synchronischen Wortbildungslehre. In Herbert Ernst Brekle and Dieter Kastovsky (eds.), *Perspektiven der Wortbildungsforschung*. Bonn: Bouvier. 155–164.
- Lipka, Leonhard. 2002. *English lexicology*. Tübingen: Narr.
- Longman dictionary of contemporary English (LDOCE)*. 2009. 5th edn with CD-ROM. Harlow: Pearson.
- Longman pronunciation dictionary (LPD)*. Wells, John C. 2008. Harlow: Pearson.
- Lowe, Solomon. 1737. *English grammar reformd* [. . .]. London.
- Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.
- Lyons, John. 1977. *Semantics*. Volume I. Cambridge: Cambridge University Press.
- Macmillan English dictionary for advanced learners (MED)*. 2007. 2nd edn. with CD-ROM. Oxford: Macmillan.
- Mair, Christian. 2009. *Twentieth-century English: History, variation and standardization*. Cambridge: Cambridge University Press.
- Marchand, Hans. 1960a. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. Wiesbaden: Harrassowitz.
- Marchand, Hans. 1960b. Die Länge englischer Komposita und die entsprechenden Verhältnisse im Deutschen. *Anglia* 78. 411–416.
- Marchand, Hans. 1969. *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. 2nd edn. Munich: Beck.
- Matthews, Peter. 1974. *Morphology: An introduction to the theory of word-structure*. Cambridge: Cambridge University Press.

- McDermott, John. 1990. *Punctuation for now*. London: Macmillan.
- McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.
- McQueen, James M. and Anne Cutler. 1998. Morphology in word recognition. In Andrew Spencer and Arnold M. Zwicky (eds.), *The handbook of morphology*. Oxford: Blackwell. 406–427.
- Meibauer, Jörg. 2003. Phrasenkomposita zwischen Wortsyntax und Lexikon. *Zeitschrift für Sprachwissenschaft* 22(2). 153–188.
- Merriam-Webster. 2001. *Merriam-Webster's guide to punctuation and style*. 2nd edn. Springfield, MA: Merriam-Webster.
- Meyer, Charles F. 1987. *A linguistic study of American punctuation*. Frankfurt am Main: Peter Lang.
- Meyer, Paul Georg, Andreas Frühwirth, Birgit Haupt, Elmar Kerz, Andreas Kohn, Timo Lothmann, Peter Marsden and Tanja Oelkers. 2005. *Synchronic English linguistics*. 3rd edn. Tübingen: Narr.
- Milroy, James and Lesley Milroy. 1985. *Authority in language: Investigating language prescription and standardisation*. London: Routledge and Kegan Paul.
- Modiano, Marko. 2002. Standardization processes and the Mid-Atlantic English paradigm. In Andrew R. Linn and Nicola McLelland (eds.), *Standardization: Studies from the Germanic languages*. Amsterdam: Benjamins. 229–252.
- Mondorf, Britta. 2000. Wider-ranging vs. more old-fashioned views on comparative formation in adjectival compounds/derivatives. In Bernhard Reitz and Sigrid Rieuwerts (eds.), *Proceedings of the Anglistentag 1999, Mainz*. Trier: Wissenschaftlicher Verlag Trier. 35–44.
- Mondorf, Britta. 2003. Support for *more*-support. In Günter Rohdenburg and Britta Mondorf (eds.), *Determinants of grammatical variation in English*. Berlin: de Gruyter. 251–304.
- Mondorf, Britta. 2009. How lexicalization reflected in hyphenation affects variation and word-formation. In Andreas Dufter, Jürg Fleischer and Guido Seiler (eds.), *Describing and modeling variation in grammar*. Berlin: Mouton de Gruyter. 361–388.
- Moon, Rosamund. 1997. Vocabulary connections: Multi-word items in English. In Norbert Schmitt and Michael J. McCarthy (eds.), *Vocabulary: Description, acquisition, and pedagogy*. Cambridge: Cambridge University Press. 40–63.
- Moore, David S. and William I. Notz. 2006. *Statistics: Concepts and controversies*. New York: W. H. Freeman.
- Morton, John. 1969. Interaction of information in word recognition. *Psychological Review* 76(2). 165–178.
- Morton Ball, Alice. 1939. *Compounding in the English language: A comparative review of variant authorities with a rational system for general use and a comprehensive alphabetic list of compound words*. New York: Wilson.
- Morton Ball, Alice. 1951. *The compounding and hyphenation of English words*. New York: Funk and Wagnalls.
- Munske, Horst Haider. 1983. Zur Fremdheit und Vertrautheit der 'Fremdwörter' im Deutschen: Eine interferenzlinguistische Skizze. In Dietmar Peschel (ed.),

- Germanistik in Erlangen: Hundert Jahre nach der Gründung des deutschen Seminars*. Erlangen: Universitätsbund Erlangen-Nürnberg. 559–593.
- Murphy, Lynne. 2010. *Lexical meaning*. Cambridge: Cambridge University Press.
- Murray, Janet H. 1997. *Hamlet on the holodeck: The future of narrative in cyberspace*. New York: The Free Press.
- Nebbrig, Alexander and Carlos Spoerhase. 2012. Für eine Stilistik der Interpunktion. In Alexander Nebbrig and Carlos Spoerhase (eds.), *Die Poesie der Zeichensetzung: Studien zur Stilistik der Interpunktion*. Berlin: Peter Lang. 11–31.
- Neijt, Anneke. 2002. The interfaces of writing and grammar. In Martin Neef, Anneke Neijt and Richard Sproat (eds.), *The relation of writing to spoken language*. Tübingen: Niemeyer. 11–34.
- Nevalainen, Terttu. 2015. The predictive potential of the S-curve model of change in diachronic studies. In Christina Sanchez-Stockhammer (ed.), *Can we predict linguistic change? Studies in variation, contacts and change in English*. Helsinki: VARIENG. www.helsinki.fi/varieng/series/volumes/16/nevalainen/. (17 August 2017)
- New English–Irish Dictionary* (NEID). www.focloir.ie. (17 August 2017)
- Newman, John and Sally Rice. 2006. Transitivity schemas of English EAT and DRINK in the BNC. In Stefan Th. Gries and Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter. 225–260.
- Nunberg, Geoffrey. 1990. *The Linguistics of punctuation*. Menlo Park, CA: Center for the Study of Language and Information.
- Okada, Takeshi. 2005. A corpus-based study of spelling errors of Japanese EFL writers with reference to errors occurring in word-initial and word-final positions. In Vivian James Cook and Benedetta Bassetti (eds.), *Second language writing systems*. Clevedon: Multilingual Matters. 164–183.
- Osselton, Noel E. 2005. Hyphenated compounds in Johnson's Dictionary. In Jack Lynch and Anne McDermott (eds.), *Anniversary essays on Johnson's Dictionary*. Cambridge: Cambridge University Press. 160–174.
- Oxford advanced learner's dictionary* (OALD). 2005. 7th edn. with CD-ROM. Oxford: Oxford University Press.
- OED: *Oxford English dictionary* (OED). 2009. 2nd edn. on CD-ROM. Version 4.0.0.3. Oxford: Oxford University Press.
- Palmer, Frank Robert. 1981. *Semantics*. 2nd edn. Cambridge: Cambridge University Press.
- Parkes, Malcolm B. 1992. *Pause and effect: An introduction to the history of punctuation in the West*. Aldershot: Ashgate.
- Partridge, Eric. 1953. *You have a point there: A guide to punctuation and its allies*. London: Hamish Hamilton.
- Peters, Pam. 2004. *The Cambridge guide to English usage*. Cambridge: Cambridge University Press.
- Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press.

- Plag, Ingo. 2010. Compound stress assignment by analogy: The constituent family bias. *Zeitschrift für Sprachwissenschaft* 29(2). 243–282.
- Plag, Ingo, Gero Kunter and Sabine Lappe. 2007. Testing hypotheses about compound stress assignment in English: A corpus-based investigation. *Corpus Linguistics and Linguistic Theory* 3(2). 199–233.
- Plag, Ingo, Gero Kunter, Sabine Lappe and Maria Braun. 2008. The role of semantics, argument structure, and lexicalization in compound stress assignment in English. *Language* 84(4). 760–794.
- Procter, Paul. 1982. *Longman new universal dictionary*. London: Longman.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
- Rabinovitch, Simon. 2007. Thousands of hyphens perish as English marches on. London: Reuters. 21 September 2007. www.reuters.com/article/2007/09/21/us-britain-hyphen-idUSHAR15384620070921. (17 August 2017)
- Rakić, Stanimir. 2009. Some observations on the structure, type frequencies and spelling of English compounds. *SKASE Journal of Theoretical Linguistics* 6. www.skase.sk/Volumes/JTL13/pdf_doc/04.pdf. (17 August 2017)
- Rakić, Stanimir. 2010. Some further observations on the spelling of English compounds. *ELOPE* 7. www.sdass.edu.si/Elope/PDF/ElopeVol7Rakic.pdf. (17 August 2017)
- Rauch, Irmengard. 1989. Language change in progress: Privacy and ‘firstness’. In Irmengard Rauch and Gerald F. Carr (eds.), *The semiotic bridge: Trends from California*. Berlin: Mouton de Gruyter. 375–384.
- Realí, Florencia and Thomas L. Griffiths. 2009. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition* 111. 317–328.
- Reiser, Karen. 2007. To hyphenate or not to hyphenate: Knowing when to hyphenate compound words. http://copyediting-grammar-style.suite101.com/article.cfm/to_hyphenate_or_not_to_hyphenate. (19 November 2007) <http://archive.li/CHWVK>. (17 August 2017)
- Ritter, Robert M. 2005a. *New Hart’s rules: The handbook of style for writers and editors*. Oxford: Oxford University Press.
- Ritter, Robert M. (ed.). 2005b. *New Oxford dictionary for writers and editors*. 2nd edn. Oxford: Oxford University Press.
- Roach, Peter. 2000. *English phonetics and phonology: A practical course*. 3rd edn. Cambridge: Cambridge University Press.
- Roach, Peter, James Hartman and Jane Setter. 2006. *English pronouncing dictionary*. 17th edn. Cambridge: Cambridge University Press.
- Rohdenburg, Günter. 2003. Cognitive complexity and *horror aequi* as factors determining the use of interrogative clause linkers in English. In Günter Rohdenburg and Britta Mondorf (eds.), *Determinants of grammatical variation in English*. Berlin: de Gruyter. 205–249.
- Rosch, Eleanor. 1973. On the internal structure of perceptual and semantic categories. In Timothy E. Moore (ed.), *Cognitive development and the acquisition of language*. New York: Academic Press. 111–144.

- Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology, General* 104(3). 192–233.
- Rovai, Francesco. 2012. On some Latin morphological (ir)regularities. In Thomas Stolz, Hitomi Otsuka, Aina Urdze and Johan van der Auwera (eds.), *Irregularity in morphology (and beyond)*. Berlin: Akademie. 187–209.
- Sanchez, Christina. 2008. *Consociation and dissociation: An empirical study of word-family integration in English and German*. Tübingen: Narr.
- Sanchez-Stockhammer, Christina (ed.). 2015. *Can we predict linguistic change? (Studies in Variation, Contacts and Change in English 16.)* Helsinki: VARIENG. www.helsinki.fi/varieng/series/volumes/16/. (17 August 2017)
- Sanchez-Stockhammer, Christina. 2017. Copy and write: The transformative power of copying in language. In Corinna Forberg and Philipp W. Stockhammer (eds.), *The transformative power of the copy: A transcultural and interdisciplinary approach*. Heidelberg: Heidelberg University Publishing. 127–148. DOI <http://dx.doi.org/10.17885/heup.195.260>. (17 August 2017)
- Sandra, Dominiek. 1990. Processing and representational aspects of compound words in visual word recognition: An experimental approach and a methodological appraisal. Amsterdam: PhD thesis.
- Sauer, Hans. 1985. Die Darstellung von Komposita in altenglischen Wörterbüchern: Studies in Mem. of Angus Cameron. In Alfred Bammesberger (ed.), *Problems of Old English lexicography*. Regensburg: Pustet. 267–315.
- Saussure, Ferdinand de. 1916/1959. *Course in general linguistics*. Transl. by Wade Baskin. New York: Philosophical Library.
- Schler, Jonathan, Moshe Koppel, Shlomo Argamon and James Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*. http://u.cs.biu.ac.il/~schlerj/schler_springsympo6.pdf. (17 August 2017)
- Schmid, Hans-Jörg. 2008. New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia* 126 (1). 1–36.
- Schmid, Hans-Jörg. 2011. *English morphology and word-formation: An introduction*. Berlin: Schmidt.
- Schmitt, Norbert. 2000. *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Schneider, Edgar W. 1997. Language change: The state of the art. In Uwe Böker and Hans Sauer (eds.), *Anglistentag 1996 Dresden: Proceedings*. Trier: Wissenschaftlicher Verlag Trier. 49–60.
- Schönefeld, Doris. 2006. From conceptualization to linguistic expression: Where languages diversify. In Stefan Th. Gries and Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter. 297–344.
- Scragg, Donald G. 1974. *A history of English spelling*. Manchester: Manchester University Press.

- Sebba, Mark. 2007. *Spelling and society. The culture and politics of orthography around the world*. Cambridge: Cambridge University Press.
- Seidenberg, Mark S. 1987. Sublexical structures in visual word recognition: Access units or orthographic redundancy? In Max Coltheart (ed.), *Attention and performance XII. The psychology of reading*. Hillsdale, NJ: Erlbaum. 245–264.
- Sepp, Mary. 2006. *Phonological constraints and free variation in compounding: A corpus study of English and Estonian noun compounds*. City University of New York: PhD dissertation.
- Shorter Oxford English dictionary* (SOED). 2007. 6th edn. Oxford: Oxford University Press.
- Simon, John. 1980. *Paradigms lost: Reflections on literacy and its decline*. London: Penguin.
- Skousen, Royal. 2002. Issues in analogical modeling. In Royal Skousen, Deryl Lonsdale and Dilworth S. Parkinson (eds.), *Analogical modeling: An exemplar-based approach to language*. Amsterdam: Benjamins. 27–48.
- Smith, Barry (ed.). 1988. *Foundations of Gestalt theory*. Munich: Philosophia.
- Snider, Neal. 2009. Similarity and structural priming. *Cognitive Science Society* 31. 815–820.
- Soanes, Catherine. 2011. Hyphens in the headlines. <http://blog.oxforddictionaries.com/2011/10/hyphens-in-the-headlines/>. (17 August 2017)
- Stang, Christian. 1993. Der Bindestrich: Regelausweitung, heutiger Gebrauch, Reformbestrebungen. *Deutsch als Fremdsprache: Zeitschrift zur Theorie und Praxis des Deutschunterrichts für Ausländer* 30(3). 163–167.
- Stefanowitsch, Anatol. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2(1). 61–77.
- Stein, Gabriele. 1974. Word-formation and language teaching. *Die Neueren Sprachen* 23(4). 316–331.
- Stein, Gabriele. 1985. Word-formation in modern English dictionaries. In Robert Ilson (ed.), *Dictionaries, lexicography and language learning*. Oxford: Pergamon. 35–44.
- Strumpf, Michael and Auriel Douglas. 1988. *Webster's new world guide to punctuation*. New York: Prentice Hall.
- Strunk, William and Elwyn B. White. 2000. *The elements of style*. 4th edn. London: Longman.
- Sundby, Bertil. 1997. The description of compounds in early English grammars: Studies in honour of Wolfgang Viereck on the occasion of his 60th birthday. In Heinrich Ramisch and Kenneth Wynne (eds.), *Language in time and space*. Stuttgart: Steiner. 221–232.
- Suttle, Laura K. and Adele E. Goldberg. Submitted. Learning what not to say: A comparison of conservatism via entrenchment and statistical preemption in adults. www.princeton.edu/~adele/Media/Pubs-by-topic_files/Learning%20what%20not%20to%20say-submitted.pdf. (17 August 2017)
- Swan, Michael. 2005. *Practical English usage*. 3rd edn. Oxford: Oxford University Press.

- Taft, Marcus. 2004. Morphological decomposition and the reverse base frequency effects. *The Quarterly Journal of Experimental Psychology* 57A(4). 745–765.
- Tagg, Caroline. 2009. A corpus linguistics study of text messaging. Birmingham: PhD thesis.
- Tagliamonte, Sali and Harald R. Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
- Tavosanis, Mirko. 2007. A causal classification of orthography errors in web texts. *Proceedings of AND 2007*. 99–106. http://research.ihost.com/and2007/cd/Proceedings_files/p99.pdf. (17 August 2017)
- Taylor, John R. 1989. *Linguistic categorization: Prototypes in linguistic theory*. Oxford: Clarendon.
- Taylor, John R. 2005. *The mental corpus: How language is represented in the mind*. Oxford: Oxford University Press.
- Terasawa, Jun. 1994. *Nominal compounds in Old English: A metrical approach*. Copenhagen: Rosenkilde and Bagger.
- Tieken-Boon van Ostade, Ingrid. 2010. The usage guide: Its birth and popularity. *English Today* 26(2). 14–23.
- Tieken-Boon van Ostade, Ingrid. 2012. Codifying the English language. In Anne Schröder, Ulrich Busse & Ralf Schneider (eds.), *Codification, canons and curricula: Description and prescription in language and literature*. Bielefeld: Aisthesis. 61–77.
- Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Tournier, Jean. 1985. *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Paris: Champion-Slatkine.
- Truss, Lynne. 2003. *Eats, shoots & leaves: The zero tolerance approach to punctuation*. London: Profile Books.
- Tulloch, Sara. 1991. *The Oxford dictionary of new words*. Oxford: Oxford University Press.
- Ungerer, Friedrich, Gerhard E. H. Meier, Klaus Schäfer and Shirley B. Lechler. 1984. *A grammar of present-day English*. Stuttgart: Klett.
- Vallins, George Henry. 1954. *Spelling*. London: Deutsch.
- Vannier, Patrick. 2013. L'Académie française: Perspectives historiques et actualités. Paper held at the University of Erlangen-Nuremberg, 15 January, 2013.
- Venezky, Richard L. 1967. English orthography: Its graphical structure and its relation to sound. *Reading Research Quarterly* 2(3). 75–105.
- Venezky, Richard L. 1970. *The structure of English orthography*. The Hague: Mouton.
- Venezky, Richard L. 1999. *The American way of spelling: The structure and origins of American English orthography*. New York: Guilford Press.
- Vollhardt, Kurt Peter C. and Neil E. Schore. 2011. *Organische Chemie*. Transl. by Katrin-M. Roy. Weinheim: Wiley-VCH.
- Waite, Maurice. 1995. *New Oxford spelling dictionary*. 2nd edn. Oxford: Oxford University Press.

- Wetzel, Claus-Dieter. 1981. *Die Worttrennung am Zeilenende in altenglischen Handschriften*. Frankfurt am Main: Peter Lang.
- Wiese, Richard. 2004. How to optimize orthography. *Written Language and Literacy* 7(2). 305–331.
- Wilbers, Stephen. 1997. Spelling compounds with and without hyphens. www.wilbers.com/part224.htm. (19 November 2007)
- Winters, Margaret E. 1990. Towards a theory of syntactic prototypes. In Savas L. Tsohatzidis (ed.), *Meanings and prototypes: Studies in linguistic categorization*. London: Routledge. 285–306.
- Wittgenstein, Ludwig. 1972. *Philosophical investigations*. Transl. by Gertrude E. M. Anscombe. Oxford: Blackwell.

CompText Corpus Texts

Newspaper Articles

- Hall, John. 2013. Revealed: Eerie new images show forgotten French apartment that was abandoned at the outbreak of World War II and left untouched for 70 years. *The Independent*. 13 May. www.independent.co.uk/news/world/europe/revealed-eerie-new-images-show-forgotten-french-apartment-that-was-abandoned-at-the-outbreak-of-world-war-ii-and-left-untouched-for-70-years-8613867.html.
- Haydon, Harry and Felix Allen. 2013. Chris Huhne and Vicky Pryce released early from prison. *The Sun*. 13 May. www.thesun.co.uk/sol/homepage/news/4925497/chris-huhne-and-vicky-pryce-released-early-from-prison.html.
- Marsden, Sam and Victoria Ward. 2013. Tia Sharp murder: Stuart Hazell changes his plea to guilty. *The Telegraph*. 13 May. www.telegraph.co.uk/news/uknews/crime/10053483/Tia-Sharp-murder-Stuart-Hazell-changes-his-plea-to-guilty.html.
- Watt, Nicholas. 2013. David Cameron rebukes ministers for saying they would vote to leave EU. *The Guardian*. 13 May. www.guardian.co.uk/politics/2013/may/13/david-cameron-cabinet-eu-referendum/print.

Academic Articles

- Ingold, Tim. 2007. Materials against materiality. *Archaeological Dialogues* 14 (1). 1–16.
- Miller, Daniel. 2008. The uses of value. *Geoforum* 39. 1122–1132.

Nonfiction Books

- Carter, Mike. 2011. *One man and his bike: A 5,000 mile, life-changing journey round the coast of Britain*. London: Ebury Press.
- Paxman, Jeremy. 2012. *Empire*. London: Penguin.
- Winchester, Simon. 2010. *Atlantic: A vast ocean of a million stories*. London: Harper Press.

Miscellaneous

- March, Bridget. 2013. Behind the scenes at the TV BAFTAs: Celeb beauty secrets. *Cosmopolitan*. 13 May. www.cosmopolitan.co.uk/beauty-hair/news/trends/celebrity-beauty/behind-the-scenes-beauty-at-the-tv-baftas.
- McKenna, Chris. 2013. Arsenal 4 Wigan 1. *BBC Sports*. 14 May. www.bbc.co.uk/sport/o/football/22430454.
- Oliver, Jamie. 2013. Mexican street salad. 13 May. www.jamieoliver.com/recipes/vegetables-recipes/mexican-street-salad.

Index

- Académie Française, 67
academy of the English language, 66–67
acceptability, 8, 69, 276, 291, 301, 306, 310, 327, 334, 354
acronym, 28, 38, 59, 107, 108, 163, 257, 298, 300
ad hoc compound, 9, 41, 113, 309
adjective compound, 18, 151–152, 173–175, 186, 197, 263, *See* attributive vs. predicative use
affix, 153–162
algorithm, 273, 276–306, 326, 328, 332, 334, 341, *See* CompSpell algorithm
allography, 311
ambiguity, 23, 63, 68, 72, 81, 82, 88, 246, 308
analogy, 234–242, 324–328
apostrophe, 113, 238, 308, *See also* genitive compound
attestation, first, 216–218
attributive vs. predicative use, 187–192, 300

back-formation, 38, 58, 59
blocking, 142
blocking principle, 302, 303
boundary *See* constituent boundaries
British vs. American English, 3, 4, 18, 67, 78, 81, 94, 229–232, 268, 271, 341

capitalisation, 42, 89, 90, 107–108, 112, 257, 281, 295, 300, 302
change, 339–344, *See also* development of compound spelling
chaos, 2, 67, 255, 271, 275, 348, 351, *See also* consistency
clipping, 38, 46, 58, 59
cognitive perspectives, 328–338
collocation, 41–42, 58
combining form, 107, 160–162, 300, 362, 364
comma, 56, 80, 108, 373
common ideas about compound spelling, 1–4
complexity, 162–164, 254, 257–258
compound
 definition, 23–24, 57–60
 spelling variants *See* spelling variants
 types, 43–57
 vs. multi-word unit, 39–42
 vs. name, 42–43
 vs. other lexemes, 35–39
 vs. phrase, 35
 word lists *See* Master List, Master_1–4, Master_5+, Master_5+_tendency, OHS_600, OHS_extra
compound stress *See* fore-stress
CompSpell, 99, 103
CompSpell algorithm, 287, 293, 294, 304, 305, 340, 352, 354
CompSpell program, 88–95, 101, 105, 112, 115, 117, 135, 136, 147, 235, 237, 247, 250, 265
CompText corpus, 81, 85, 86, 289, 290, 305, 351, 361, 362, 392
conscious, 1, 64, 269, 275, 299, 329, 330, 332, 333, 334, 336, 353
consistency, 1, 4, 63, 65, 147, 227, 228–229, 245, 275, 336
consonant cluster, 98–100, *See* garden path cluster
constituent boundaries, 8, 53, 56, 57, 88, 102, 138, 308, 312, 324, 337
constituent family, 234–242
 frequency, 239–242
 size, 234–238
construction, 8
control, 63
conversion, 39, 44, 58, 59
coordination, 32
copulative compound, 27, 33, 45, 47
corpora, 81–88

decision tree, 277, 281, 283, 285, 288, 296, 297, 298, 326, 332, 333
derivation *See* suffix
development of compound spelling, 3, 132, 215–225, 258, 339, 342–344
dictionaries, 10–13, 77–80
doubt, 62, 72, 290, 303, 304, 321, 347

- economy, 132, 246–253
 -ed, 37, 155, 158–160
 editing, 228–229
 education, 63, 65, 66, 68–69, 71, 301
Elements of Style, 10
 elliptical compound, 55, 94, 309
 emphasis, 234
 exception principles, 299–303
 expert, linguistic, 70–71
- family resemblance, 264, 324
 foreign origin, 57, 210–215
 Germanic, 210–214
 Romance, 210–214
 fore-stress, 26–27
 Fowler, 2, 9, 107, 132, 309, 310, 348
 frequency, 132–144. *See also* constituent family frequency
- garden path cluster, 102–105, 111, 255, 260, 271
 general reference, 197–200, 209, 300, 302
 genitive compound, 45, 109–111, 155, 179
 government, 68–69, 72
GPO Style Manual, 9, 23, 57, 68, 108, 197, 205, 210, 235
 grammatical compound, 172, 176–177, 287, 373–374
 grapheme, 98, 311
 graphotactics, 97–108
- headedness, 33, 60, 168, 208, 300
 heterogeneous constituents, 256–257, 262, 263, 271, 309
 hyphen, functions of, 308–309
 hyphenated compound *See* spelling variants
- identity, 64
 idiomaticity, 206–208
 ill-formedness, 28, 45
 inflection *See* suffix
 inflection tolerance principles, 90–93
 -ing, 6, 115, 155, 156, 158–160, 168, 211, 266, 300, 302, 309, 367
 italics, 57, 210
- language user, 71
 length, 44, 112–131
 lexicalisation, 255, 258–260
 -ly, 173–175
- man*, 37, 113, 198, 200, 235
 markedness, 23, 174, 264, 309–310, 334
 Master List, 79–80, 88–89
 Master_1–4, 148, 350
- Master_5+, 97, 113, 114
 Master_5+_tendency, 279, 304
 medium, 246
 blog, 246–253
 chat, 246–249
 text message, 84, 249–253, 344
 memory, 113, 245, 324, 326, 329, 336
 mental lexicon, 33, 40, 324, 326, 331–332, 341
 minimal pair, 233–234, 312
 mistake, 61, 62, 65
 multidimensional prototype model, 319–322
- native speaker study, 291–295
 neoclassical compound, 36, 46. *See* combining form
 neologism, 220, 330
 noncanonical ordering, 160, 168
 norm, 61–73
 noun+noun compound, 5, 6, 18, 25, 56, 77, 135, 146, 158, 272, 323, 328, 345, 353
 numbers, 47, 53, 108, 176–177, 179, 257, 280, 291
- obscured compound, 36, 80
 OHS_600, 96, 356–360
 OHS_extra, 97
 open compound *See* spelling variants
 optimal compound spelling, 335–338, 353
 orthographic unity, 25–26
 orthography vs. spelling, 14
- part of speech, 47–53, 171–195
 particle, 48, 95
 permanent relationship, 26, 34, 147
 phrasal verb, 39, 59, 176, 206
 phrase compound, 33, 39, 45, 58, 59, 116, 165, 179, 227
 prefix, 157–158, 364
 prepositional verb, 39, 59, 176
 prescriptive, 8, 9, 10, 12, 14, 64, 67, 70, 73, 182, 192, 310
 probabilistic, 324, 325, 327, 331
 processing *See* cognitive perspectives
 proofreading, 55, 65
 prototype, 313–324
 prototypical compound, 323–324
 publisher, 4, 9, 66, 69–70, 85, 280, 338
 punctuation indicator, 308
 punctuation, minimal, 245
- readability obstacle, 255, 260–261
 reduplication, 205–206
 register, 227–228
 resting activation, 332, 341, 353

- sentential paraphrase, 193–195
- significant compound spelling determinants, 365–370
- simplex lexeme, 7, 35–36, 59, 337
- single-letter constituent, 119–120
- slash, 56, 80, 329, *See also* spelling variants
- solid compound *See* spelling variants
- spatial restriction, 249–253
- species-genus compound, 205
- speed of typing, 246–249, 344
- spellchecker, 15–16, 17, 65, 247, 341
- spelling variants
 - alternative spellings, 55–57
 - hyphenated compound, 54, 300, 302
 - open compound, 54, 300, 302
 - relation between main types, 307–313
 - solid compound, 54, 300, 302
- standardisation, 63, 71, 142, 230, 344, 347
- state of the art, 4–16
- stress, 281, 296, 301, 303, *See also* fore-stress
- style guides, 3, 8–10, 11, 17, 62, 66, 69, 171, 192, 235, 296, 297
- subconscious, 227, 286, 299, 329, 330, 332, 333, 347, *See also* conscious
- suffix, 37, 236, 362, *See also* complexity, *-ing*, *-ed*, *-ly*
 - inflectional, 157
 - lexical, 154
- syllabification, 114–116
- take-the-best heuristics, 284–286, 304, 351
- tradition, 63, 72
- triconstituent compound, 53, 112–114, 161, 162, 323
- unconscious, 72, *See also* conscious
- underlying clause *See* sentential paraphrase
- unified semantic concept, 35
- uninterruptability, 29–30, 176
- US Government Printing Office, 9, 68, *See also* *GPO Style Manual*
- variables
 - coded, 265–268
 - correlated, 277, 295
 - significant *See* significant compound spelling determinants
- variation, 4, 8, 61, 69, 80, 94, 141–143, 223, 298, 306, 327, 330, 333, 334, 336, 339, 349
- verb compound, 176